The challenge of connecting your server to storage

Fabrics and more

Lowell Vanderpool & Nathan Vanderpool
TECH SAVVY PRODUCTIONS

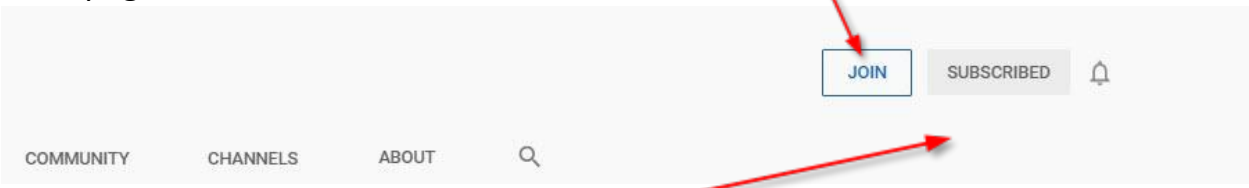
CONTACT US:

mrvanderpool@techsavvyproductions.com

nathan@techsavvyproductions.com

SUPPORT US:

If you would like to support the channel, Join our channel membership, it's $2.99/month (less than a Starbucks coffee); see the "Join" button on our channel homepage.



OR

Subscribe to the channel as it helps our channel perform better on YouTube's algorithm.

SOCIAL MEDIA AND WEBSITE:

Check out our **Website**: https://www.techsavvyproductions.com

Follow us on **Twitter:** @_TechSavvyTeam

Like us on **Facebook:**
https://www.facebook.com/Tech-Savvy-Productions-105287381500897

Mr.V **Linkedin**: https://www.linkedin.com/in/lowell-vanderpool-57970623/

Nathan **Linkedin**: _https://www.linkedin.com/in/nathan-vanderpool-50a27822/

Follow on **Instagram**: techsavvyproductions
https://www.instagram.com/techsavvyproductions/

We translate subtitles on our videos into the following languages: عربى, български, 简

Tech Savvy Productions

体中文), 中國傳統的）, Nederlands, Suomalainen, Pilipino, français, Deutsche, हिंदी, Magyar, bahasa Indonesia, 日本語, 한국어, norsk, Polskie, português, Română, русский, Española, Kiswahili, Svenska, and Tiếng Việt

Social media logos and “Tech Savvy Productions” teaser created by The 11th Hour: https://www.youtube.com/user/The11thH0UR/featured

Quick Links to topics in Notes

Q&A (Part 1) from “Storage Trends for 2021 and Beyond” Webcast

SATA vs. NVMe, Is it time for NVMe ?

Pavilion compares RocE and TCP NVMe over Fabrics performance

BLOCK STORAGE VS. OBJECT STORAGE: WHEN TO USE EACH

Block Storage Advantages and Disadvantages

WHAT IS NETWORK FILE SYSTEM (NFS)?

NAS VS. SAN VS. DAS – ADVANTAGES & DISADVANTAGES

CNA storage NICs

Server Disaggregation: Sometimes the Sum of the Parts Is Greater Than the Whole

NVMe SSD Classification

Data Center Cabling Solution: DAC Cables vs AOC Cables

What is a SAN LUN?

Solid State Drive Form Factors

Storage Spaces Direct overview

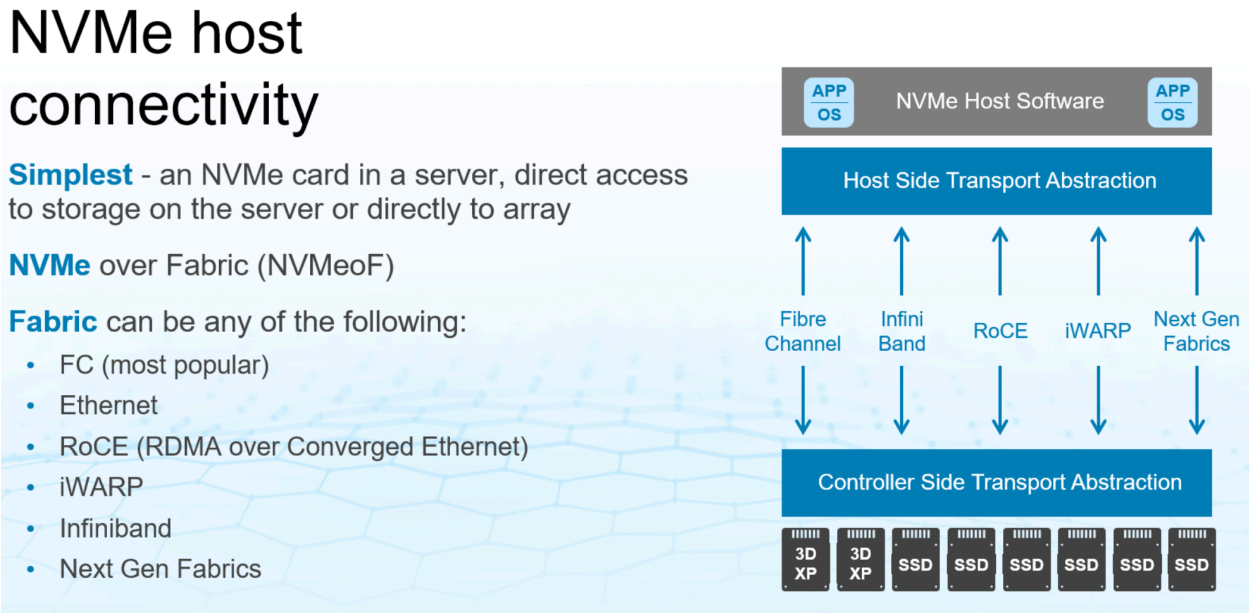
Adaptec® SmartRAID 3200 RAID Adapters new 24G SAS controllers

AFF A800 system installation and setup instructions: https://youtu.be/b6Ys24AHS7A

What is a network fabric?

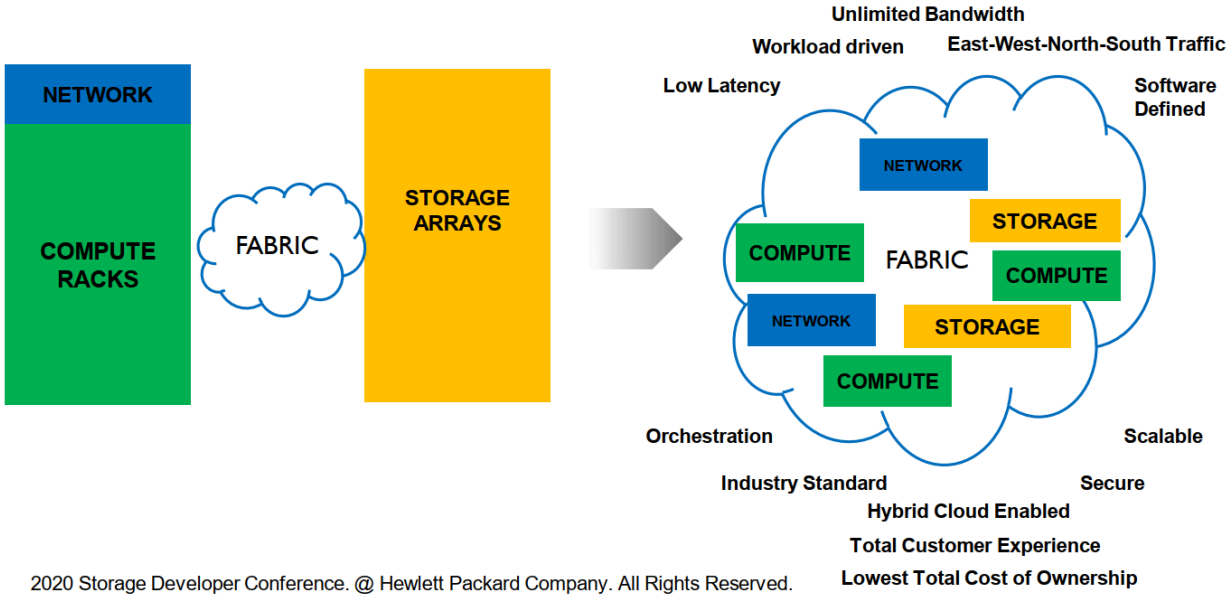
A fabric is an advanced type of network based on a switched architecture that is not based on a ring or bus type architecture. Most fabrics offer multiple paths between endpoints, a well-defined repeating architecture and a very scalable set of management tools that can support hundreds or thousands of endpoints.

The new NVMe SSD interfaced can be connected across a Fabric. In fact it can be connected across lots of different fabrics: Ethernet (3 approaches), Fibre Channel, InfiniBand, and PCIe to date.



Data Centers want to share storage readily among multiple compute nodes and be able to perform clustering, failover, and other system-wide operations at NVMe SSD speeds. NVMe over Fabrics (NVMe-oF) is the solution. This talk will describe the technology in its many forms. Describe use cases, for both Enterprise and cloud, where it is being applied. Then finish with potential future directions it is heading.

Disaggregation – What does it mean?



Disaggregated storage is a type of data storage within computer data centers. It allows compute resources within a computer server to be separated from storage resources without modifying any physical connections.[1]

A form of composable disaggregated infrastructure, disaggregated storage allows resources to be connected via a network fabric providing flexibility when upgrading, replacing, or adding individual resources.

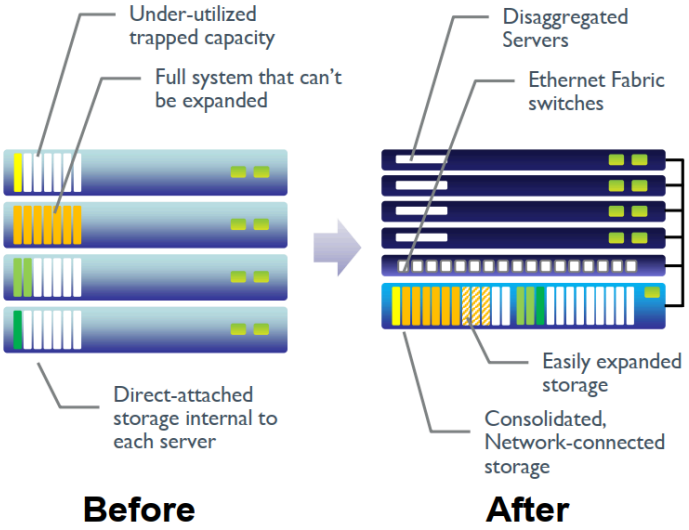
It also allows servers to be built for future growth, offering greater storage efficiency, scale and performance than traditional data storage without compromising throughput and latency.

A number of technology improvements are combining to make storage disaggregation a reality. These include:

- Modern server performance: due to the PCIe Gen 4 serial bus, many servers can deliver more than 8GB/sec of throughput, which far exceeds traditional storage networking performance capabilities.
- The shift toward NVMe: The shift from disk to SAS/SATA flash, and now NVMe flash, puts pressure on servers and networks alike. A single NVMe drive delivers millions of IOPS, far beyond the usual capabilities of conventional storage networking.[9]

- 10Gb/25Gb/40Gb/100Gb Ethernet. More and more data centers are replacing slower network connections with faster Ethernet, removing bandwidth limitations and bottlenecks.[10]

NVME-oF Use Case - Redefining Internal DAS



- Advantages:
  - Delivers the performance of DAS
  - Improves utilization of flash and facilitates data sharing
  - Increases availability of storage with HA and network connectivity
  - Reduces rack space and power requirements
  - Delivers better Total Cost of Ownership
  - Improves customer experience deploying NVMe-oF

A new language for accessing solid state media

Traditional Storage Arrays

1. Storage Controller runs SCSI
2. Front end FC/iSCSI
3. Backend SAS/SATA
4. Software Feature Rich based on SCSI

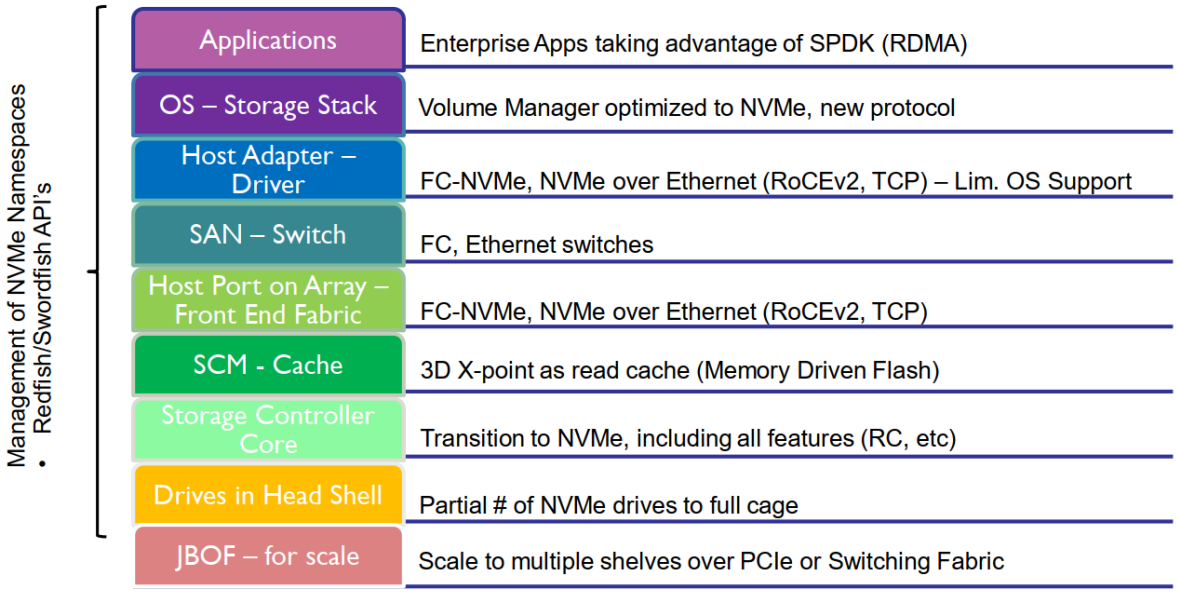
Hybrid case

1. Storage Controller runs SCSI. Upgraded back end (partial/full)– Controller does SCSI-NVMe translation with NVMe drives in the backend
2. Memory-Driven Flash
3. Software Feature Rich based on SCSI

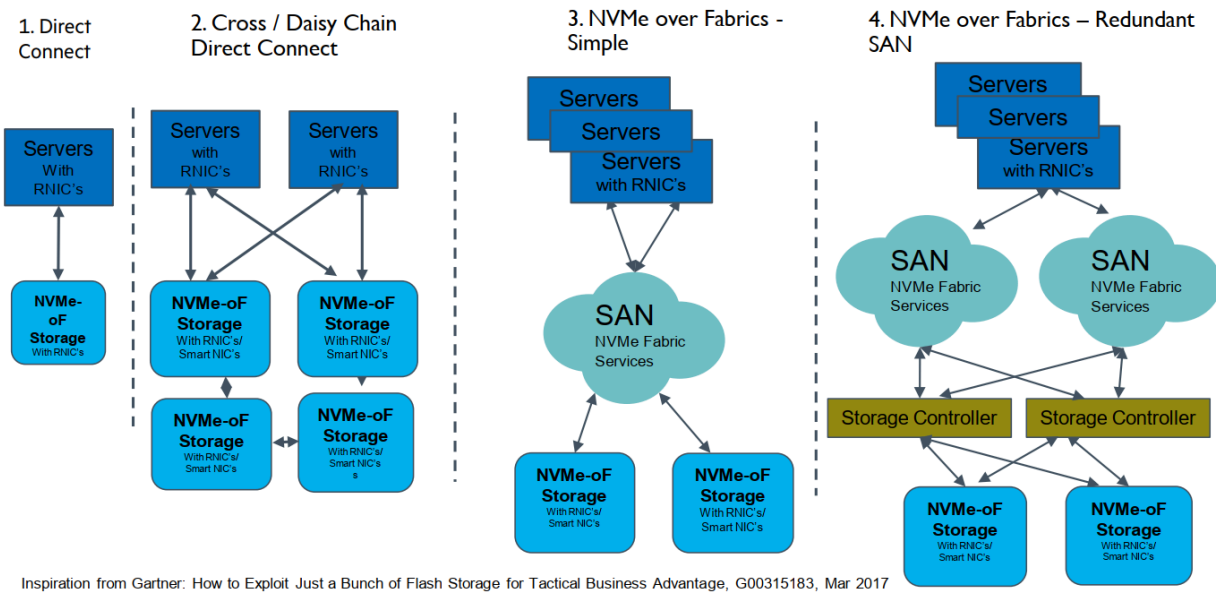
Next Gen. Storage Arrays

1. Controller runs NVMe
2. Backend NVMe Drives (PCIe, NVMe over Fabrics)
3. Frontend NVMe (FC-NVMe, NVMe over Ethernet)
4. Software Features running NVMe, expect parity in 3 years

I/O Stack evolution



NVMe over Fabrics Deployment Scenarios



Inspiration from Gartner: How to Exploit Just a Bunch of Flash Storage for Tactical Business Advantage, G00315183, Mar 2017

The landscape today….

Protocol	Latency	Scalable	Performance	Hybrid Enterprise
Fibre Channel	Lower	Yes	High	Teir 0, On-Prem
RoCEv2	Lowest	Yes	High	Tier 0, Hybrid
TCP	Low-Medium w/Offload	Yes	Medium-High	Tier 1, Hybrid
InfiniBand	Lowest	Limited	High	None
iWARP	Medium	Yes	Medium	None

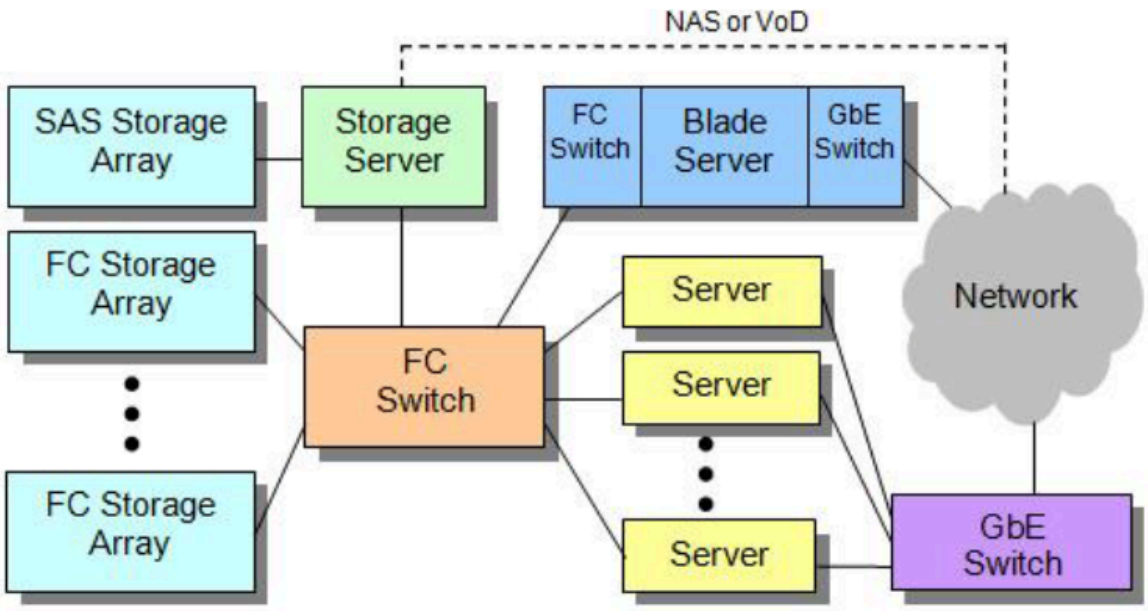
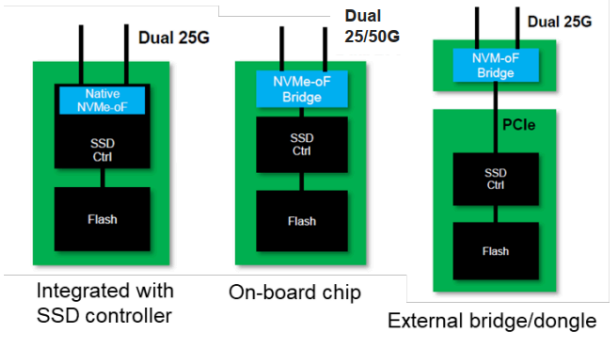


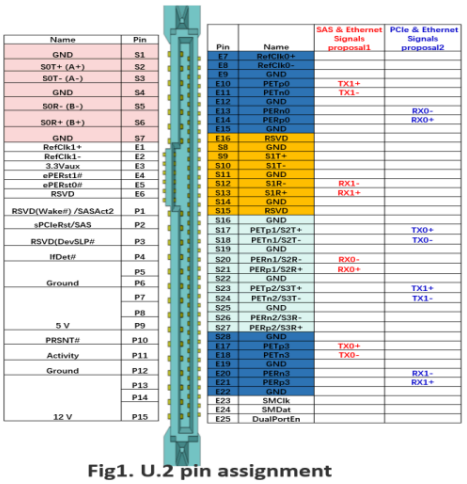
Figure 1. A Traditional Data Center Storage Network Block Diagram

Ethernet SSDs

eSSDs

- Different eSSD designs today
- Some will support multiple interfaces and protocols
  - Ethernet, PCIe, SAS, SATA
  - RoCE, TCP





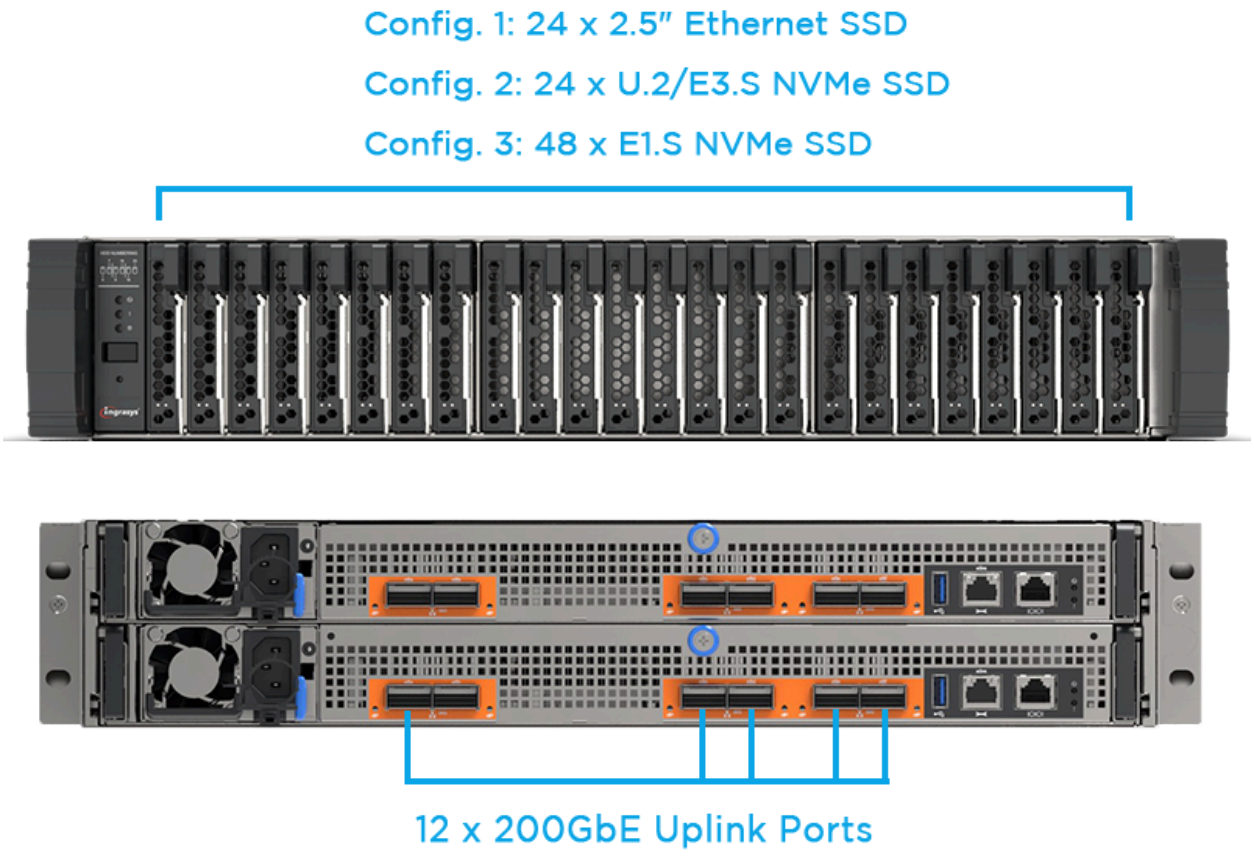
Name	Pin
GND	S1
S0T+ (A+)	S2
S0T- (A-)	S3
GND	S4
S0R- (B-)	S5
S0R+ (B+)	S6
GND	S7
RefClk1+	E1
RefClk1-	E2
3.3Vaux	E3
ePERst1#	E4
ePERut0#	E5
RSVD	E6
RSVD(Wake#) /SASAct2	P1
sPCIeRst/SAS	P2
RSVD(DevSLP#	P3
IfDet#	P4
	P5
Ground	P6
	P7
	P8
5 V	P9
PRSNT#	P10
Activity	P11
Ground	P12
	P13
	P14
12 V	P15

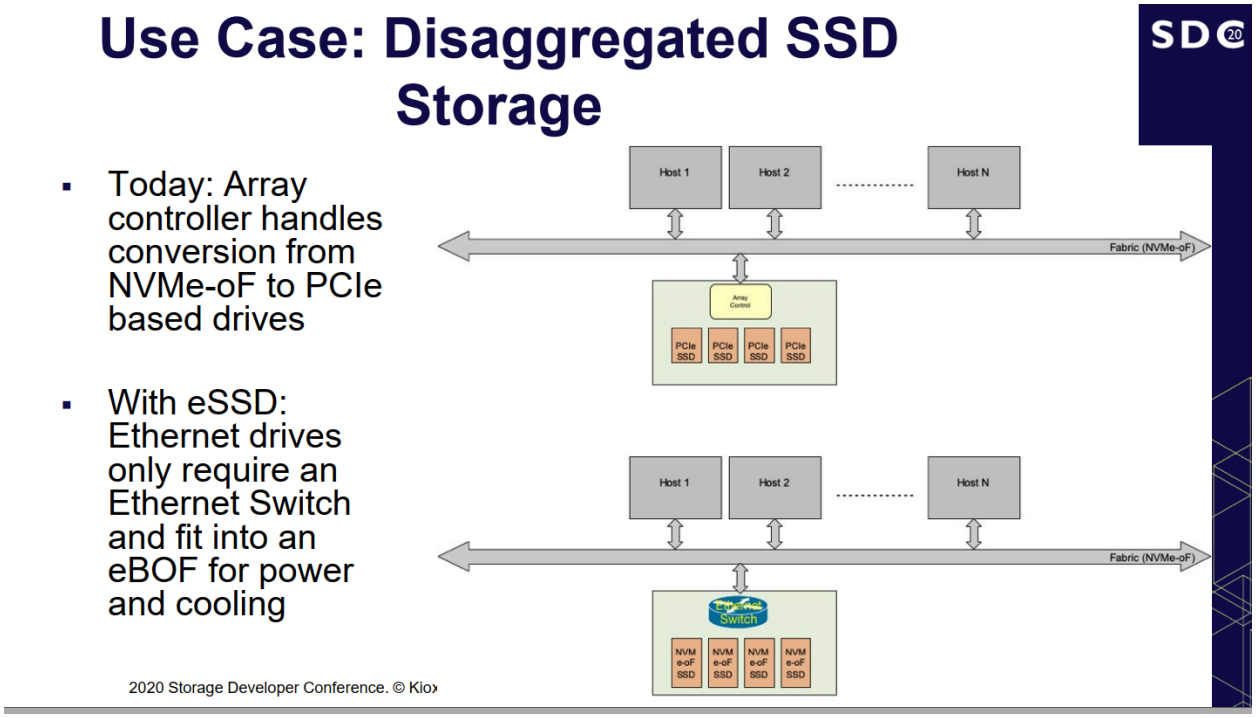
Pin	Name	SAS & Ethernet Signals proposal1	PCIe & Ethernet Signals proposal2
E7	RefClk0+		
E8	RefClk0-		
E9	GND		
E10	PETp0	TX1+	
E11	PETn0	TX1-	
E12	GND		
E13	PERn0		RX0-
E14	PERp0		RX0+
E15	GND		
E16	RSVD		
S8	GND		
S9	S1T+		
S10	S1T-		
S11	GND		
S12	S1R-	RX1-	
S13	S1R+	RX1+	
S14	GND		
S15	RSVD		
S16	GND		
S17	PETp1/S2T+		TX0+
S18	PETn1/S2T-		TX0-
S19	GND		
S20	PERn1/S2R-	RX0-	
S21	PERp1/S2R+	RX0+	
S22	GND		
S23	PETp2/S3T+		TX1+
S24	PETn2/S3T-		TX1-
S25	GND		
S26	PERn2/S3R-		
S27	PERp2/S3R+		
S28	GND		
E17	PETp3	TX0+	
E18	PETn3	TX0-	
E19	GND		
E20	PERn3		RX1-
E21	PERp3		RX1+
E22	GND		
E23	SMClk		
E24	SMDat		
E25	DualPort6n		

Fig1. U.2 pin assignment

SFF-8639 connector

SAN JOSE, CA, November 11, 2021 – KIOXIA America, Inc. today announced the production availability of its EM6 Series Enterprise NVMe-oF™ solid state drives (SSDs) for Ethernet Bunch of Flash (EBOF) systems. Using the Marvell® 88SN2400 NVMe-oF SSD converter controller that converts an NVMe® SSD into a dual-ported 25Gb NVMe-oF SSD, KIOXIA EM6 Series drives expose the entire SSD bandwidth to the network.


Config. 1: 24 x 2.5" Ethernet SSD
Config. 2: 24 x U.2/E3.S NVMe SSD
Config. 3: 48 x E1.S NVMe SSD
12 x 200GbE Uplink Ports



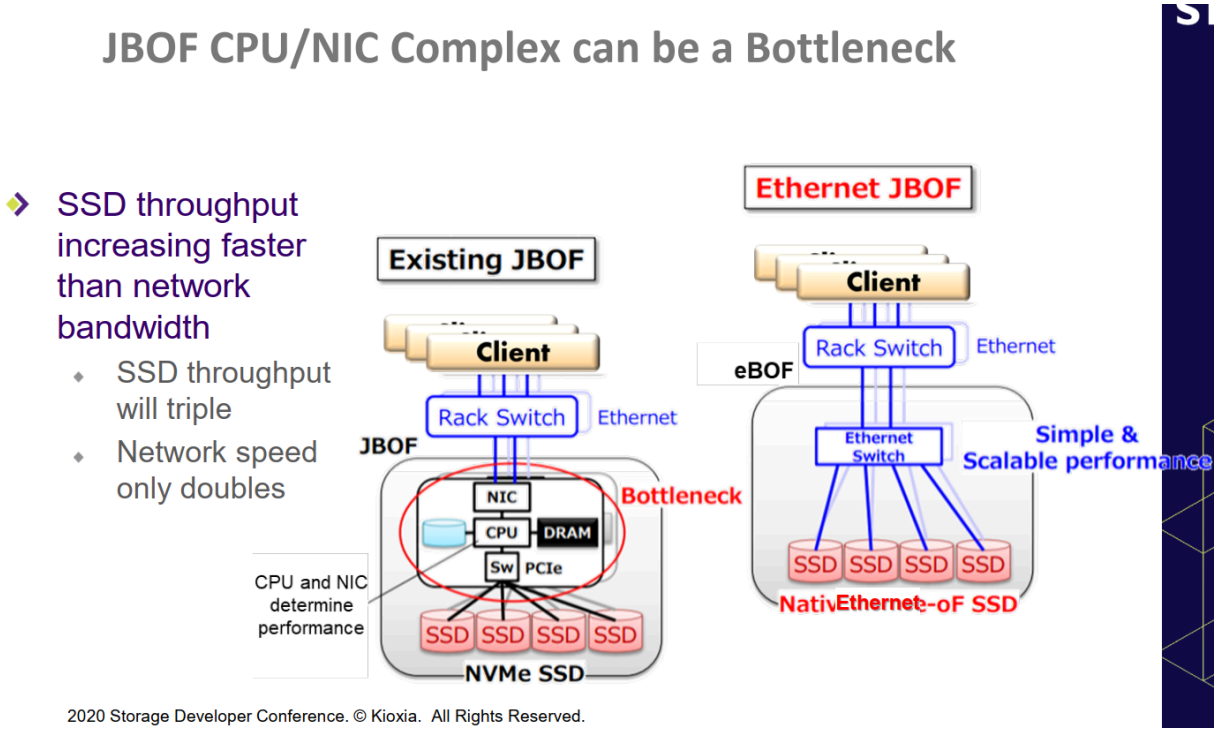
Ethernet Bunch of Flash (**EBOF**)

Ethernet Bunch of Flash in an NVMe-oF™ Network for Low-Cost Storage at Scale

By Joe Steinmetz - 2020-11-10

https://www.micron.com/about/blog/2020/november/ethernet-bunch-of-flash-in-an-nvme-of-network-for-low-cost-storage-at-scale

NVMe over Fabrics (aka NVMe-oF™) is an emerging technology that enables disaggregation of compute and storage in the data center. Micron is helping to unlock the benefits of NVMe-oF by collaborating with key technology partners such as Marvell, Foxconn-Ingrasys and NVIDIA. We are also innovating new technologies such as heterogenous-memory storage engine (HSE) that will help optimize access to flash storage at scale using NVMe-oF.



What is NVMe-oF?

NVMe-oF literally extends the NVMe protocol over a network, increasing the reach well beyond the server chassis that confines SSDs today. While NVMe has been around since 2011, the fabrics extension was first standardized in 2016. Because NVMe-oF leverages NVMe, it inherits all the benefits: a lightweight and efficient command set, multicore awareness and protocol parallelism. NVMe-oF is truly network agnostic as it supports all common fabrics, including Fibre Channel, InfiniBand and Ethernet.

Figure 1 compares NVMe and NVMe-oF models and highlights the various network and network transport options that are available to the user.

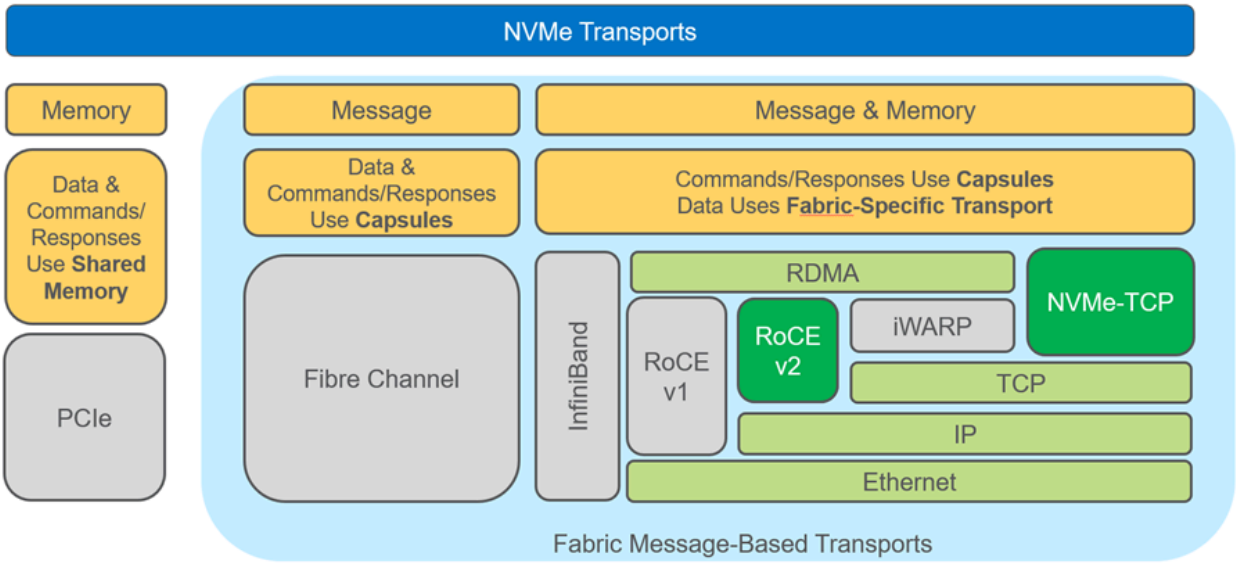


Figure 1: NVMe and NVMe-oF™ Model Comparison

There are two relevant Ethernet transport options, RoCE v2 and NVMe-TCP. Each has its advantages and disadvantages. RoCE v2 is lower latency but requires specialized RDMA-enabled NICs (RNIC), while NVMe-TCP transport has higher latency and higher CPU use but does not require any specialized RNICs. Instead, it makes use of a standard NIC. RoCE v2 is more prevalent in the market right now.

Universal RDMA

Remote direct-memory access(RDMA) is a technology that has been in use for more than a decade. In server connectivity, data copying is a major source of processing overhead. In a conventional networking stack, received packets are stored in the operating system's memory and later copied to application memory. This copying consumes CPU cycles and also introduces latency. Network adapters that implement RDMA enable writing data directly into application memory. Applications that transfer large blocks of data, such as networked storage and virtual machine migration, reap the greatest efficiency gains from RDMA.

What are the benefits of NVMe over Fabrics?

With just NVMe, you are essentially restricted to the server chassis or the rack using PCIe switches as a means of scaling. While this is a perfectly valid way of scaling storage, it is arguably limited in scope and reach. NVMe-oF

allows a virtually unlimited amount of storage to be connected across a data centerwide radius.

Today, NVMe-oF is well established, with many users embracing the technology to connect all-flash arrays (AFAs) to servers. However, the full benefit of NVMe-oF will only be realized when compute and storage are fully disaggregated. That is, a pool of NVMe SSDs is made available over the network to a pool of servers in a way that allows one to provision both compute and storage on demand. Disaggregation increases scalability and shareability of storage and enables composability, as shown in Figure 2.

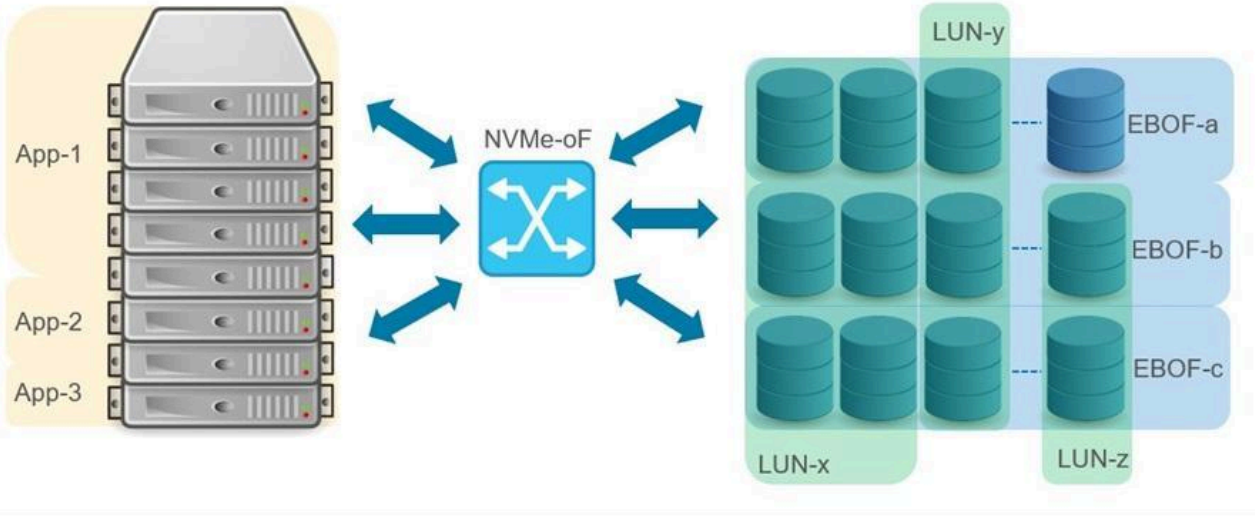


Figure 2: Disaggregation of Compute and Storage

Another dimension to disaggregated storage is storage services (that is, data protection, replication, compression, and others). Storage services can be managed by the servers (onload model) or offloaded to data processing units (DPUs) that are close to the actual storage. Tradeoffs must be made. The onload model consumes additional CPU cycles and network bandwidth but minimizes cost, while the offload model increases cost and, depending on provisioning, can create bottlenecks. The pursuit of low-cost storage at scale leads to an onload attached storage strategy due to the TCO (total cost of ownership) advantages.

What are EBOFs, JBOFs and JBODs?

There are two ways to connect a “bunch of flash” into a NVMe-oF network: using an Ethernet Bunch of Flash (EBOF) or using a Just a Bunch of Flash (JBOF). Don’t confuse a JBOF with a JBOD (Just a Bunch of Disks). A JBOD is typically used to scale storage in a rack using NVMe over PCIe. An EBOF or JBOF can be used to scale storage across a data center using NVMe-oF. As seen in Figure 3, a JBOF uses a PCIe switch to fan out to the SSDs, while the EBOF uses an Ethernet switch to fan out to the SSDs. Both a JBOF and an EBOF connect back to the servers using NVMe-oF.

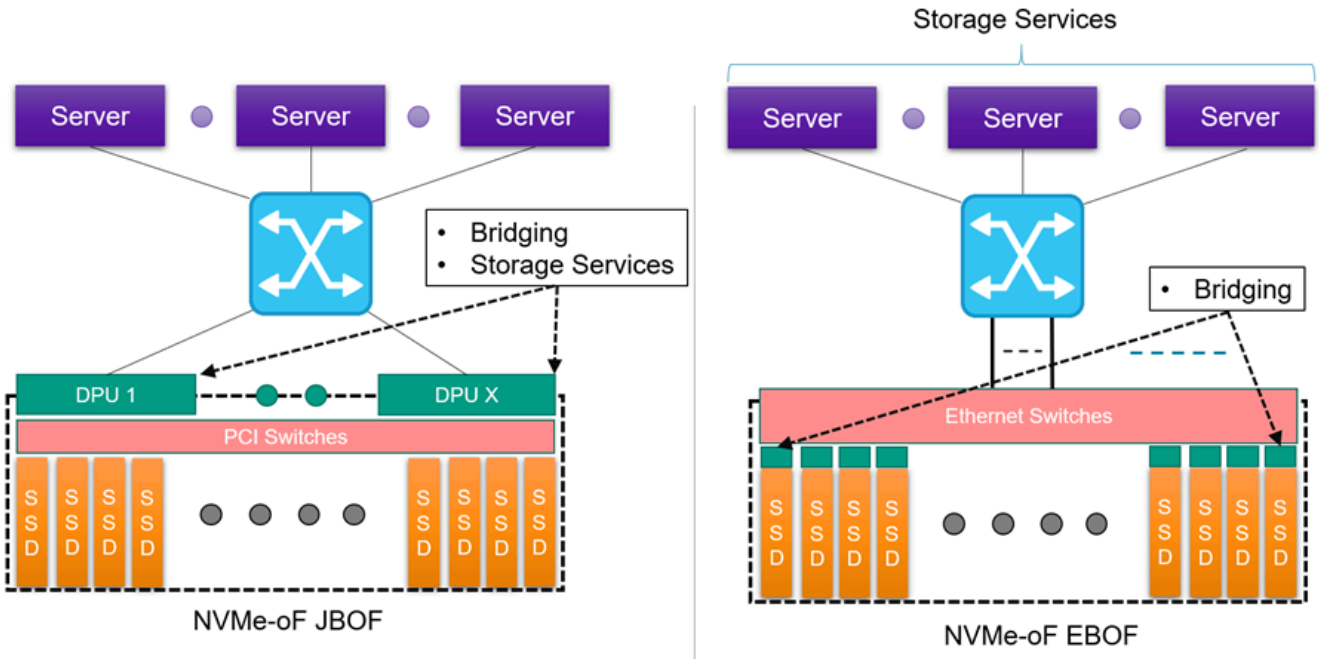


Figure 3: EBOF and JBOF Comparison

The main difference between the two approaches, beyond the obvious Ethernet vs. PCIe switching, is where the NVMe to NVMe-oF conversion takes place. On the JBOF, the conversion or bridging is at the periphery of the shelf using one or more DPUs (x DPUs to y SSDs, x:y ratio). On the EBOF, the bridging is done within the SSD carrier or enclosure (x bridges to x SSDs, 1:1 ratio). While the JBOF has an advantage of using the processing capabilities of the DPU for running storage services, it does present a potential bottleneck and comes at additional cost and power over the EBOF model. The cost tradeoff and bottlenecking come into play when the ratio of bridges to SSDs is not 1:1.

We're testing our system with the Marvell 88SN2400 and Foxconn-Ingrasys EBOF

Through a collaboration with Marvell and Foxconn-Ingrasys, we've been testing our Micron 7300 mainstream NVMe SSDs in NVMe-oF environments under a variety of different applications and workloads.

Before looking at this testing, let's look at the Foxconn-Ingrasys EBOF and Marvell's 88SN2400 converter controller and Prestera® CX 8500 switch.

Marvell's 88SN2400 is an NVMe-oF SSD converter controller for cloud and enterprise data centers. This, in combination with the Marvell switch, essentially allows you to convert or "bridge" between NVMe and NVMe-oF. The 88SN2400 converter controller is a critical component to the Foxconn-Ingrasys EBOF and, together with our Micron 7300 SSDs, makes for an impressive high-performance 2U shelf of storage (up to 73.1 GB/s of bandwidth and up to 20 million IOPs). Figure 4 shows the Foxconn-Ingrasys EBOF, with 24 U.2 slots in a 2U enclosure.

Figure 4: Foxconn-Ingrasys EBOF

Figure 5 displays the Foxconn-Ingrasys SSD carrier with the Marvell 88SN2400 converter controller.

Figure 5: Foxconn-Ingrasys U.2 Carrier With 88SN2400

The Foxconn-Ingrasys U.2 carrier takes a standard U.2 SSD form factor. The U.2 carrier supports dual Ethernet ports to address applications that need path redundancy, and it has a single PCIe Gen3 x4 on the drive side for the NVMe SSD.

Marvell's 88SN2400 converter controller supports both RoCE v2 and NVMe-TCP transports. However, for our testing, we've focused on RoCE v2.

How do things scale with NVIDIA™ GPUDirect™ Storage (GDS)?

We've been doing a lot of work with our SSDs in artificial intelligence and machine learning workloads using NVIDIA™ GPUDirect™ Storage (GDS). We

wanted to see how things scaled in a fabric environment by connecting a Foxconn-Ingrasys EBOF with Marvell's 88SN2400 converter controller to a NVIDIA DGX™ A100 system. This is a simple gdsio (GPUDirect Storage I/O) tool test comparing bandwidth and latency both with and without GDS in an NVMe-oF environment.

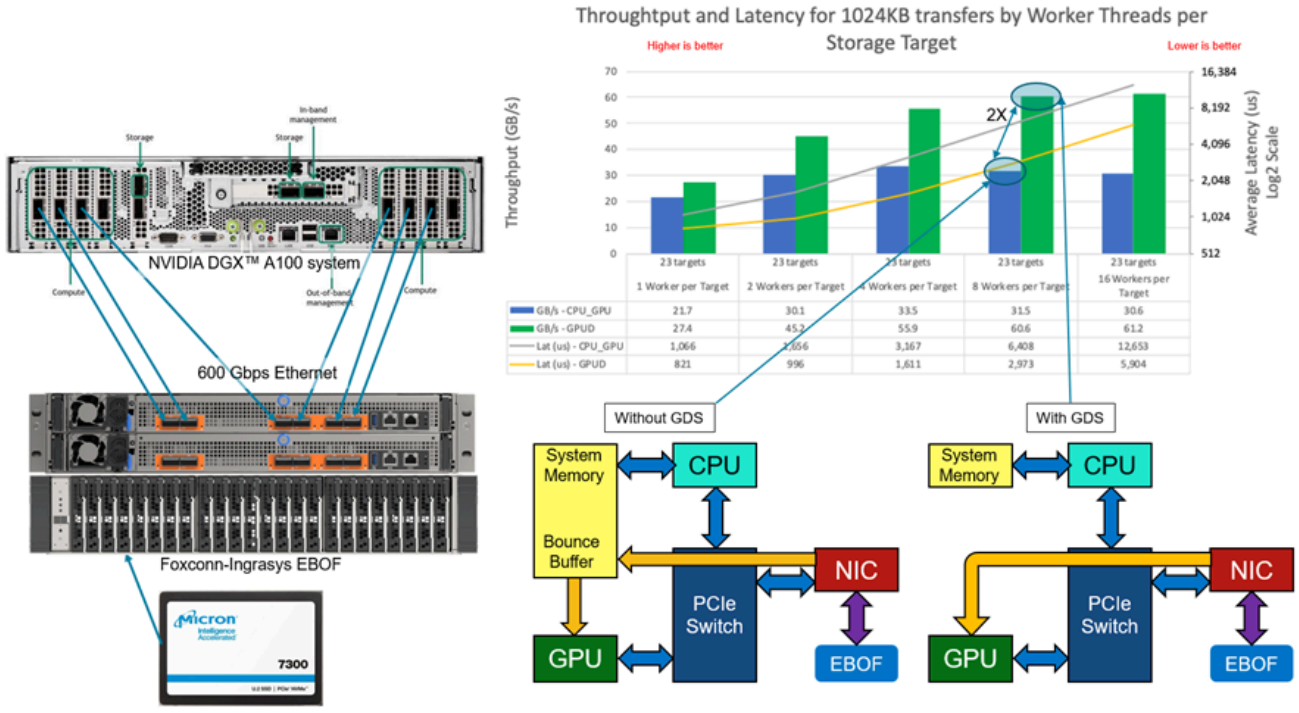


Figure 6: DGX™ A100 with EBOF

In Figure 6, we have an EBOF loaded with Micron 7300 SSDs connected directly to an NVIDIA DGX™ A100 using six of the eight compute network ports, giving 600 Gb/s of available network bandwidth. GDS enables data to be transferred directly between peers, skipping the high-latency bounce buffer that is used when GDS is not enabled. In this test, we are extracting the full capabilities of the SSDs in aggregate (~61 GB/s) for the workload. Future testing will add an Ethernet switch and scale up the number of EBOFs even more.

You can learn more about this testing at FMS 2020 in the AI track via a presentation by Wes Vaske, principle storage solutions engineer, entitled "Analyzing the Effects of Storage on AI Workloads."

How can NVMe-oF create scale with an HSE?

Here at Micron, we've been working on some amazing technologies, one of them being the heterogeneous-memory storage engine (HSE). HSE is a flash-aware storage engine that enhances the performance of storage-class memory (SCM) and SSDs. It also increases the effective SSD life span through reduced write amplification, all while being deployed at a massive scale. NVMe-oF is an ideal way to further create scale with HSE. To validate the effectiveness of HSE in the context of fabric attached storage, we've done some testing using MongoDB with YCSB (Yahoo! Cloud Serving Benchmark). In Figure 7, we compare performance between the default built-in MongoDB storage engine (WiredTiger) using local Micron 7300 SSDs and Micron's HSE using Micron 7300 SSDs in an EBOF.

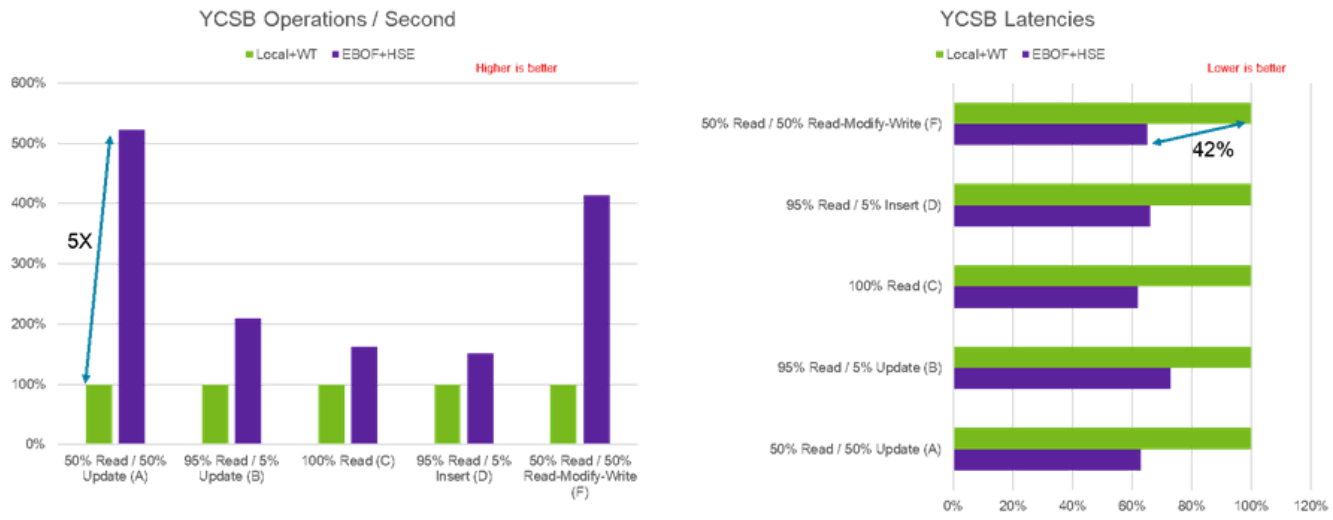


Figure 7: WiredTiger compared to HSE

The effectiveness of HSE in a fabric environment is quite dramatic when compared to the legacy WiredTiger storage engine used in MongoDB with a local SSD. We can achieve up to five times the improvement in YCSB operations per second and a 42% reduction in latency while simultaneously increasing the scalability of storage.

You can learn more about this testing at FMS 2020 in a presentation made by Sujit Somandepalli, principal storage solutions engineer, entitled "Extend Your Storage With NVMe Over Fabrics."

What is the future of NVMe-oF?

NVMe-oF is an enabling technology that will eventually lead to fully disaggregated data centers where applications can be composed and then dynamically provisioned with the appropriate amount of compute and storage in a cost-effective manner.

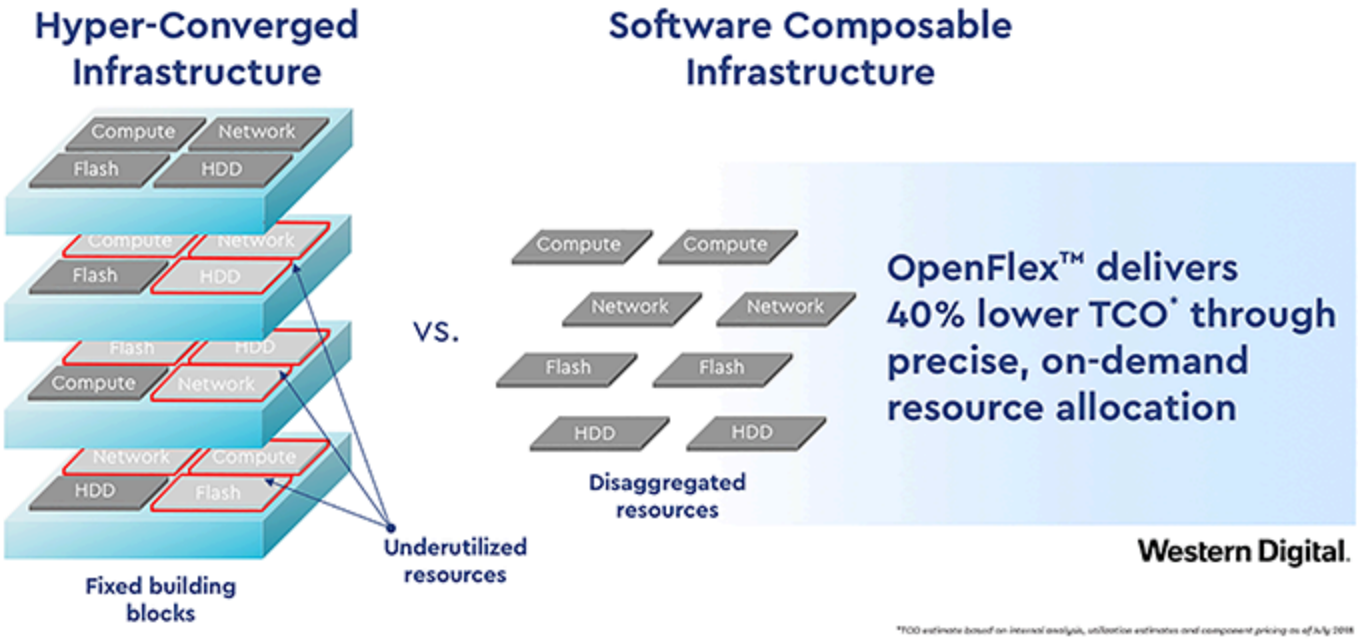
Today, low-cost bridges or DPU-based platforms are used to connect and bridge NVMe SSDs into an EBOF or JBOF. In the future, we may see native NVMe-oF SSDs further decreasing TCO and improving performance.

Micron is designing next-generation data center SSDs with capabilities and features that are optimized for NVMe-oF applications.

Western Digital Announces OpenFlex Storage Architecture and NVMeoF Storage Devices

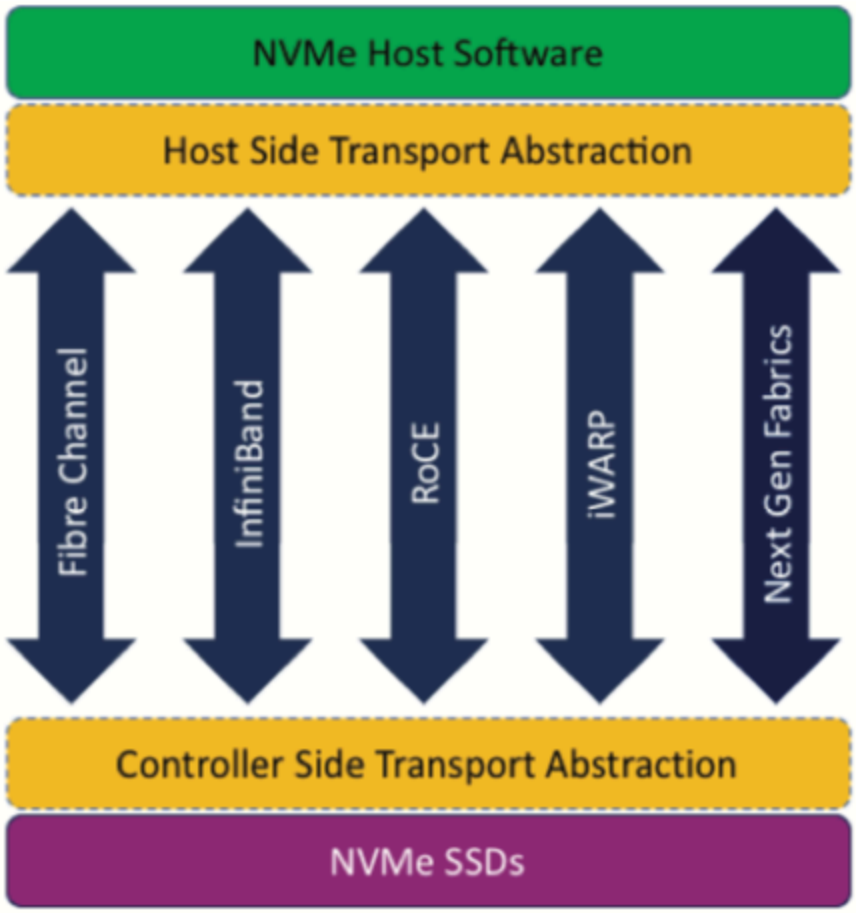
by Anton Shilov *on August 7, 2018 3:00 PM EST*

https://www.anandtech.com/show/13180/western-digital-announces-openflex-architecture-and-nvmeof-storage-devices



Western Digital on Tuesday introduced its new OpenFlex storage architecture and the first family of products supporting it. The OpenFlex architecture promises to enable operators of datacenters to independently scale compute, storage and network resources by using software composable infrastructure (SCI). Meanwhile Western Digital will share the mechanical specifications of OpenFlex products as well as publicly release its Kingfish API for managing SCI to make OpenFlex an open industry standard.

The OpenFlex architecture relies on storage devices featuring SSDs and HDDs that use an NVMe-over-Fabric (NVMeoF) interface, and can be used to create independently scalable pools of storage connected to computing resources using standard technologies (such as Ethernet, InfiniBand, etc.). Western Digital says that independently scalable pools of various resources will allow customers to better utilize their installed hardware and software, therefore reducing the initial infrastructure investment and the total cost of ownership by eliminating problems like “stranded storage.”



NVMf extends NVMe for use on network fabrics

In addition to its hardware OpenFlex architecture, Western Digital is also introducing its Kingfish API, which enables pools of devices to be presented as SCI and arranged into logical application servers. Western Digital says that since said storage pools are directly connected to other resources, logical

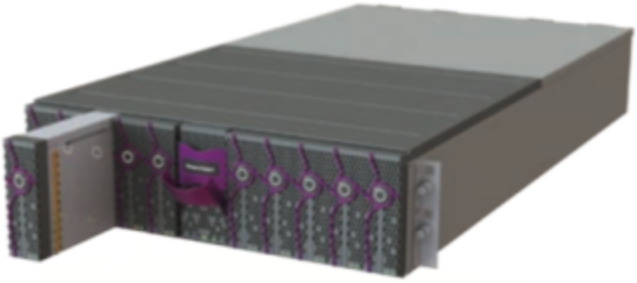
application servers will not compete for resources and therefore will have more predictable performance. The latter claim looks rather promising, but certainly needs an independent verification. Furthermore, to take advantage of this capability, the Kingfish API has to be supported by popular datacenter software.

So far Western Digital's OpenFlex has gained support from various datacenter hardware and software products as well as from companies, including Apache Hadoop, Apache Spark, Apache Kafka, Apache Cassandra, Apache Mesos, Broadcom, Ceph, DriveScale, Hewlett Packard Enterprise (HPE), Inspur, Kaminario, Kubernetes, Marvell Technology Group, Mellanox Technologies, Microsoft SQL Server, Percona, and Super Micro Computer. Technically, all of the aforementioned should support the Kingfish API as well.

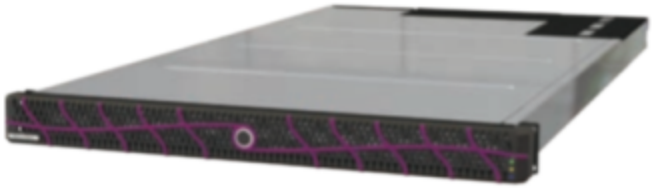
Obviously, Western Digital will be the first to offer products based on the OpenFlex architecture. Initially, there will be three devices:

- *The **OpenFlex F3000** high-performance hot-swappable solid-state storage device with two 50 GbE ports and capacities ranging from 12.8 TB to 61.4 TB.*
- *The **OpenFlex E3000** 3U enclosure that that accommodates up to 10 F3000 devices with 128 TB – 614 TB capacity.*
- *The **OpenFlex D3000** 1U enclosure with two 25 GbE ports that houses several hard drives featuring up to 168 TB capacity.*

The solid-state OpenFlex storage products will be available in Q4 2018, whereas the HDD-based D3000 1U box will launch in 2019. Though do note that Western Digital has yet to public prices for the new hardware.



OpenFlex F3000 Series fabric device and 3U enclosure for mission critical apps and data



OpenFlex D3000 Series fabric device and 1U enclosure for data tiering, data protection and disaster recovery

The Advantages of NVMe SSD

1. Lower Latency (Compared with SAS SSD)

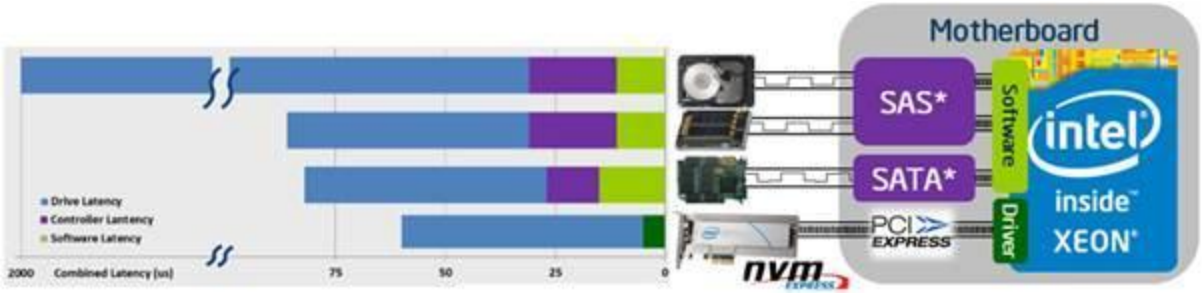


Figure1. How data are written into drive

- Green stands for Software Latency
- Purple stands for Controller Latency
- Blue stands for Drive Latency

The NVMe SSD relies on the native PCI-e controller connected to the CPU directly, instead of the traditional method which transfers through the Southbridge controller and then connects to the CPU. This streamlines the calling method. The CPU does not need to read registers when executing commands, thereby reducing latency and improving performance;

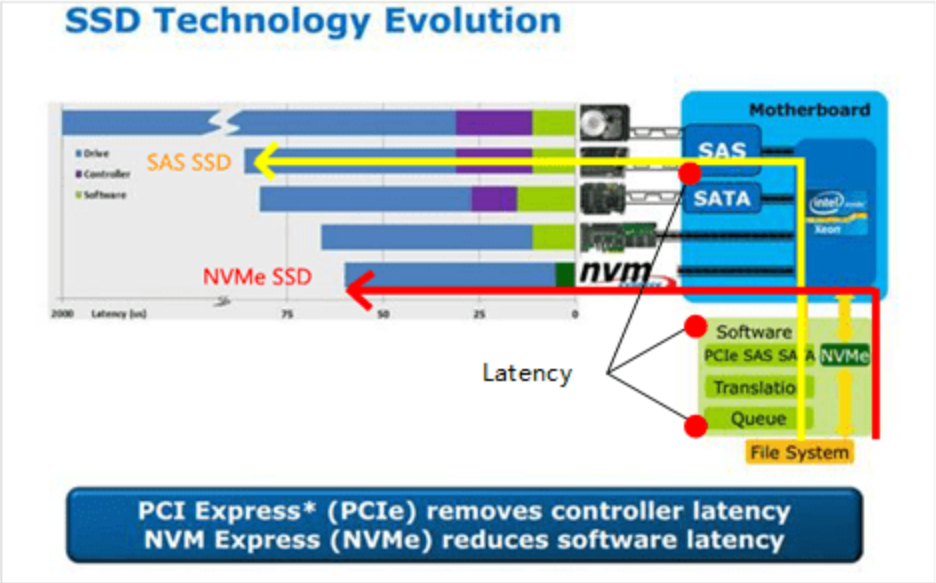


Image from communities.intel.com.

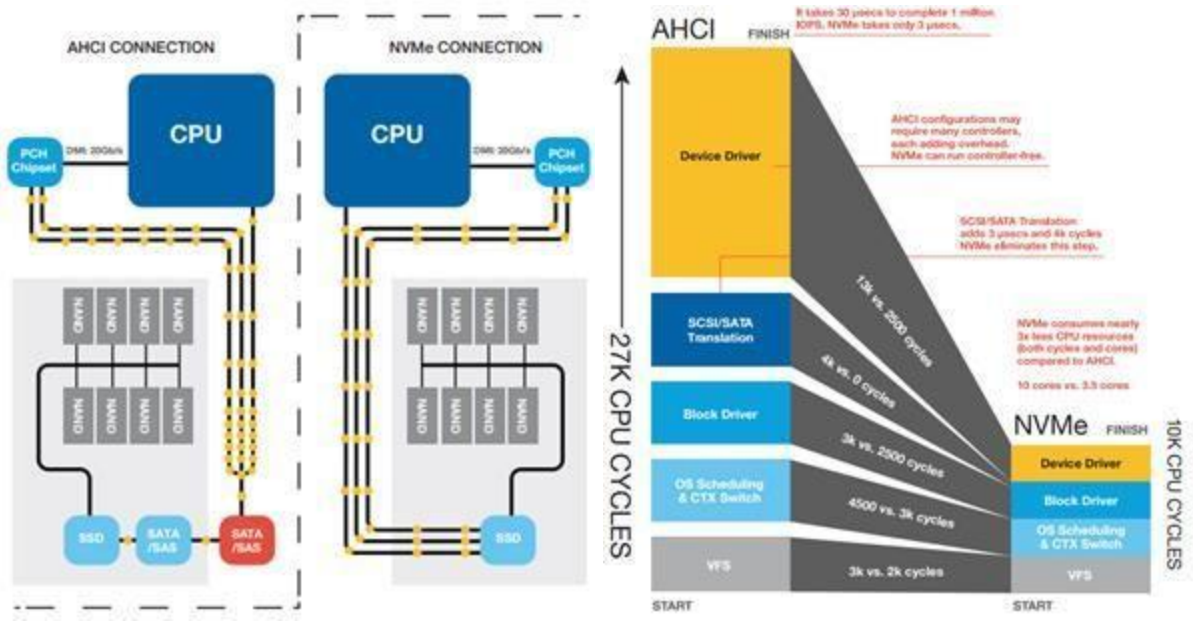


Figure 2. AHCI vs. NVMe: What it takes to reach 1 million IOPS

NVMe SSDs require fewer CPU cycles to complete the same number of IOPS and use approximately 33% less CPU resources than their AHCI/SATA counterparts.

2. NVMe SSD has Better Performance than SAS SSD

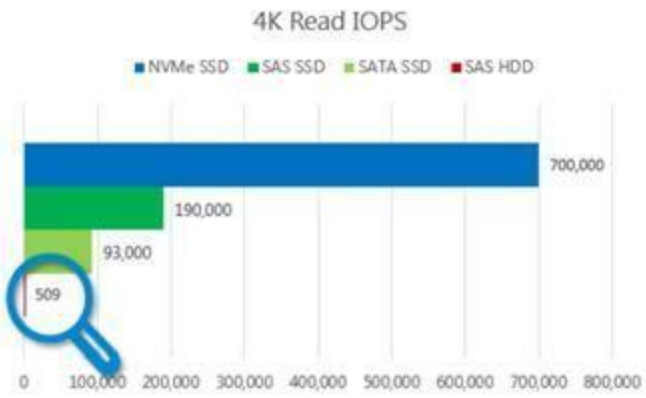


Figure 3: 4K Read IOPS by Drive Type

Figure 3 comes from Micron® Cooperation. It shows an example of 4KiB random read IOPS among three high performance Micron® SSDs (9100 PRO NVMe, S610DC SAS and 5100 ECO SATA) along with a performance SAS HDD (15,000 RPM).

This raw speed advantage can translate to noticeable performance gains in I/O intensive workloads that require fast responses such as those running on cloud and data center servers, video-on-demand, and virtualization applications, among others.

In Huawei's datasheet, with the same 7.68TB capacity, the IOPS performance gap between NVMe SSD and SAS SSD has doubled.

Figure 3: 4K Read IOPS by Huawei HSSD. On the left, is SAS SSD, the other side is NVMe SSD.

Model	Huawei HSSD V5 SAS SSD			
Specifications				
Interface specifications	SAS 3.0, SAS 12Gb/s			
Product Model	HSSD-D6223AL960N	HSSD-D6223AL1T9N	HSSD-D6223AL3T8N	HSSD-D6223AL7T6N
Available capacity	960G	1920G	3840G	7680G
Dimensions	100.2 mm x 69.85 mm x 14.7 mm			
Weight	<= 350 g			
Performance				
Average read duration (@4 KB, QD1)	135 µs			
Average write duration (@4 KB, QD1)	40 µs			
Sequential read rate	2200MBps	2200MBps	2200MBps	2200MBps
Sequential write rate	900MBps	1600MBps	1600MBps	1600MBps
Random read IOPS (@4 KB)	335K	400K	400K	400K
Random write IOPS (@4 KB)	38K	75K	100K	80K

Model	Huawei HSSD V5 NVMe SSD			
Specifications				
Product Model	HSSD-D6D23DL960N	HSSD-D6D23DL1T9N	HSSD-D6D23DL3T8N	HSSD-D6D23DL7T6N
Available capacity	960G	1920G	3840G	7680G
Interface specifications	PCIe 3.0			
Command Set	NVMe 1.2			
Dual-port Access	Dual-port			
Dimensions	100.2 mm x 69.85 mm x 14.7 mm			
Weight	<= 350 g			
Performance				
Sequential Read Bandwidth (128 KB, QD32)	3200MBps	3200MBps	3200MBps	3200MBps
Sequential Write Bandwidth (128 KB, QD32)	900MBps	1780MBps	2600MBps	2600MBps
Random Read IOPS (4 KB, QD128)	370K	700K	800K	800K
Random Write IOPS (4 KB, QD128)	37K	75K	110K	100K

3. NVMe SSD is the mainstream in the future of storage field

As developments in the SSD space continue to push and surpass the limits of AHCI technologies, NVMe over PCIe is looking more and more like a winner. Unlike AHCI, which was developed with HDD technology in mind, NVMe was designed specifically as an interface protocol for flash. SSDs that use NVMe over IP protocol are able to make full use of the advantages that flash technology has to offer over standard storage

technologies. Nowadays, there is more and more storage vendors focus on NVMe all flash array system.

Summary

The difference among NVMe、SAS、SATA、PCI-e…..

Physical Interface/Connector	Transport Layer Protocol	Host Controller Interface(Bottom of Application Layer)	Command Set (Upper Application Layer)
ATA PCMCIA CompactFlash	ATA	IDE/ATA Interface	ATA Command Set
SATA M.2 SFF8639(U.2) mSATA eSATA	SATA	ACHI	ATA Command Set
SCSI FC SAS Ethernet InfiniBand	SCSI FC SAS TCP(over IP) RDMA oE/oIB) TCP(over IP)	SCSI	SCSI Command Set
PCI-e Standard M.2 SFF8639(U.2)	PCI-e	NVMe	NVMe Command Set
eMMC/UFS	eMMC/UFS	eMMC/UFS	eMMC/UFS

That's all, thanks!

Q&A (Part 1) from “Storage Trends for 2021 and Beyond” Webcast

24G SAS »

https://www.scsita.org/library/qa-part-1-from-storage-trends-for-2021-and-beyond-webcast/

Questions from “Storage Trends for 2021 and Beyond” Webcast Answered

It was a great pleasure for Rick Kutcipal, board director, SCSI Trade Association (STA), to welcome Jeff Janukowicz, Research vice president at IDC and Chris Preimesberger, former editor-in-chief of eWeek, in a roundtable talk to discuss prominent data storage technologies shaping the market. If you missed this webcast titled “Storage Trends for 2021 and Beyond,” it’s now available on demand here.

The well-attended event generated a lot of questions! So many in fact, we’re authoring a two-part blog series with the answers. In part one, we recap the questions that were asked and answered during the webcast, but since we ran out of time to answer them all, please watch for part two when we tackle the rest.

Q1. How far along is 24G in development?
A1. Rick: The specification is done and most of the major players are investing in it today. Products have been announced and we’re also expecting to see server shipments in 2022. STA has a plugfest scheduled for July 6, 2021. It’s a busy time and everybody’s pretty excited about it!

Q2. What’s after 24G SAS?
A2. Rick: Naturally, one would think it would be a 48G speed bump, but it’s not clear that’s necessary. There’s still a lot of room for innovation within the SCSI stack, not just in the physical layer. The physical layer is the one that people can relate to and think “oh, it’s faster.” Keep in mind that there are a lot of features and functionality that can be added on top of that physical layer. The layered architecture of the SCSI stack, enables changes whether it’s at the protocol layer or another higher layer, without impacting the physical layer. These are happening real time and STA is having T10 technical committee meetings on a regular basis, and innovations are in the works.

Q3. Where does NVMe HDD and 25G ethernet HDD fit in?
A3. Jeff: Generally speaking, it’s still unclear how that’s going to evolve. As we look out over time, in the enterprise market on the SSD side, clearly, we’re seeing NVMe move into the majority of the shipments and SSDs are growing as a percentage of the overall unit shipments and petabytes. However, right now we’re seeing a mix of technologies that are used within a storage array or in an enterprise system. And clearly, they are SAS-based SSDs and HDDs. And with that transition to more SSDs, it’s sort of a natural question to say, “hey, what about

putting the NVMe interface on HDDs?" Now you obviously don't necessarily need it for all the performance reasons or the optimizations around non-volatile media, which is why NVMe was introduced, but there are some initiatives, and these could help bring some cost savings and further system optimizations to the industry. There are some things underway from OCP in terms of looking at NVMe based HDDs, but they're still relatively early on at least from my perspective in terms of their development. But there are definitely some activities underway that are looking at the technology.
Rick: From my perspective, I'm seeing a surge in NVMe HDD work within OCP. My concern with NVMe HDDs is the amount of standards work that still has to be done to make them work in an enterprise environment. I think people forget it's not just taking some media and putting an NVMe interface in front of it. How do all the drive inquiries get mapped to NVMe? How do you manage enterprise large scale spin up? I think it's an exciting time. I think there are a lot of good possibilities, but the amount of work that's needed can be underestimated sometimes.

Q4. Could you discuss the adoption of SAS SATA and NVMe in all flash arrays?
A4. Jeff: IDC has seen a lot of investment in terms of all flash arrays. And we've seen pretty rapid growth over the last couple years. In 2020, about 40% of the spending on external storage was on all flash arrays. And the reality is if you look at that today, the vast majority of those are really still built upon SAS-based SSDs. There have been some announcements from a lot of the large storage providers around NVMe-based arrays, whether it's Dell EMC, Netapp, Pure Storage, IBM, etc. Today, these solutions have already started to become available in the market. And we do see NVMe AFA's as a very high growth category over the next few years, but right now they're still targeted primarily at a lot of the higher end and more performance-oriented types of applications. We're really just starting to see them move down into the more mainstream portion of the all flash array market. Which from IDC's perspective, if it was 40% last year, we see it growing as an overall category to about 50% of the overall spend on external storage by 2023. So clearly there is a lot going on in this market as well.
Rick: My questions in regards to NVMe and all flash arrays is always about scalability. I know there's a lot of work going on regarding NVMe over fabrics, but if you go back and look at the amount of computational resources, memory and system resources that it takes to scale these things, there's still some pretty big challenges ahead. I'm not saying it's not going to happen, but of course the ecosystem, has solved hard problems in the past.

Q5. How do you differentiate between M.2 SSDs and NVMe in client system deployments?

A5. Rick: The SOCs or the controllers on these devices are very different. There are enterprise class M.2 drives, so the form factor doesn't necessarily preclude it from fitting into one of these categories. While M.2 is more designed to the client, it's not a hard and fast thing. Typically, it's the traditional 2.5.

Jeff: Rick, you're pretty much spot on. There are some differences at the SOC level and design level such as power fail protection. But there does tend to be a different firmware load a lot of times for the enterprise class drives. There can also be some differences in terms of the endurance in how those drives are designed. But if the question is about form factors, we really are at an interesting point for the industry, because historically it has always been dictated by HDD form factors. But as flash has grown, we've seen a lot of new form factors. M.2 is obviously one that was originally designed for some of the client market, and has now found its way into a lot of enterprise applications. E1 short is a slight variant of M.2 but is on the roadmap to be more enterprise optimized form factor. But we also see some other ones out there like E1 long, which is a longer version of E1.S. There's also U.3 and others which are pretty interesting in terms of ways to optimize around some of the new storage media, i.e., SSDs and solid state.

Q6. Is the NVMe takeover sooner than 3-5 years?

A6. Rick: That's a very logical question. People that aren't in the ecosystem day-to-day might not be seeing the 24G SAS adoption. Right now, there's a lot of investments at the system and sub-system level. For 24G SAS there are multiple adapter vendors, same as there has been in the past for 12G SAS. And from the media side, there are numerous drive vendors sampling 24G SAD drives today, and one has been announced. I think some people are going to be shocked of the 24G adoption, and that's going to start coming to light at STA's next plugfest, with some big demos and press announcements as products get ready to launch. So, I guess I would, say stay tuned for that one because I think people, some people, are going to be pretty surprised.

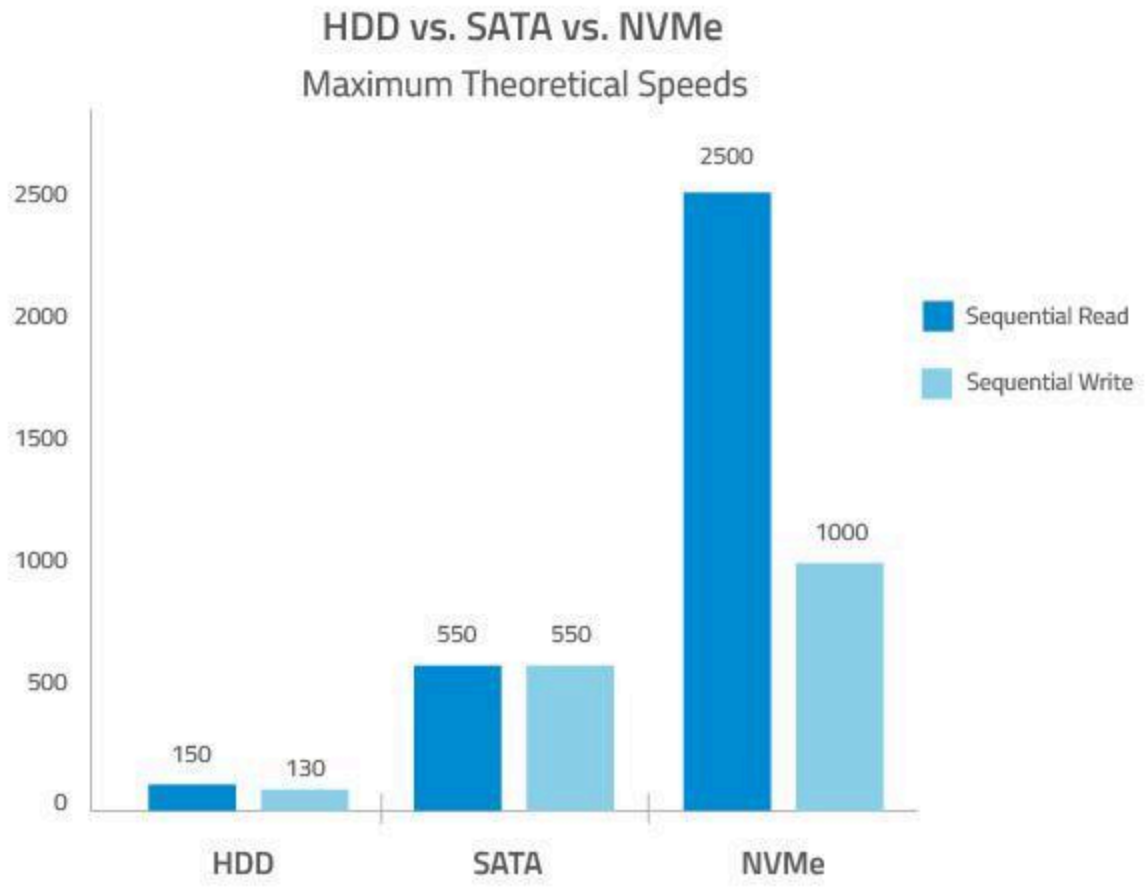
SATA vs. NVMe, Is it time for NVMe ?

SSDs2018-10-25

https://www.atpinc.com/blog/nvme-vs-sata-ssd-pcie-interface

NVMe vs. SATA: It's Time for NAND Flash in the Fast Lane

The deluge of data being created every day shows no sign of abating. As users create more and more data, we will not only need space to store it, but also higher throughput and faster response times to access data.



Over the years, CPU and DRAM technologies have been continually improving and increasing speeds to meet escalating data-hungry requirements. Mechanical storage drives such as hard disk drives and the interfaces connecting them, however, have trailed behind. Compounding the problem is that the faster, higher-capacity solid state drives (SSDs) today still widely use interfaces designed for slower mechanical drives.

Serial ATA (SATA™) is generally the least expensive and most extensively used SSD interface today. However, its latest generation is already almost a decade old, and its transfer rate is capped at 6 Gb/s. SATA uses the legacy protocol Advanced Host Controller Interface (AHCI), to connect the CPU/memory subsystem. AHCI was optimized for rotating media, resulting in high latency when used with faster storage solutions that do not have mechanical parts, such as SSDs. SATA does not support high levels of expansion, as most servers can accommodate fewer than six SATA devices.

To narrow the widening gap between fast CPU/DRAM and slow storage, Non-Volatile Memory Express or NVM Express®, also known as NVMe™, was developed by leading technology manufacturers.

NVMe™ in a Flash

NVMe™ is an interface specification developed specifically for NAND flash and next-generation SSDs. This interface leverages existing PCI Express® (PCIe®) technology to efficiently support the growing bandwidth needs of enterprise and client systems. Although PCIe basically functions as an interconnect linking motherboard-mounted peripherals like graphics or wireless network cards, its suitable high-bandwidth bus technology is ideal for today's fastest SSDs.

Flash at PCIe Speeds

One great feature of PCIe is its direct connection to the CPU. This streamlines the storage device stack, completely eliminating much of the complexity and layers present in SATA protocol stacks. As a result, NVMe delivers 2X the performance of SAS 12 Gb/s, and 4-6X of SATA 6 Gb/s in random workloads. For sequential workloads, NVMe delivers 2X the performance of SAS 12 Gb/s, and 4X of SATA 6 Gb/s.
(Source: "All About M.2 SSDs," Storage Networking Industry Association [SNIA]. 2014.)
By taking advantage of PCIe, NVMe reduces latency, enables faster access, and delivers higher Input/Output per Second (IOPS) compared with other interfaces designed for mechanical storage devices. NVMe also offers performance across multiple cores for quick access to critical data, scalability for current and future performance and support for standard security protocols.

ATP NVMe Solutions: Built for the Fast Lane

ATP Electronics responds to the clamor for acceleration in enterprise and mission-critical environments with its NVMe-based M.2 2280 SSDs. Designed for a PCIe 3.0 x4 interface and complying with NVMe 1.2 specifications, ATP NVMe SSDs boast up to 1 TB memory capacity, sequential read speed of up to 2,540 MB/s, and sequential write speeds up 1,100 MB/s. ATP NVMe solutions also integrate 3D NAND MLC technology, enabling higher memory capacity, lower cost per bit, and increased longevity.

HDD vs SATA vs. NVMe

The table below shows a comparison of HDD, SATA SSD and NVMe SSD. NVMe, as a protocol designed specifically for PCIe SSDs, delivers better performance than PCIe and SATA SSDs using the AHCI protocol.

Interface	PCIe		SATA 6 Gb/s
Protocol	NVMe Protocol (Optimized for flash SSDs)	AHCI Protocol (Optimized for mechanical HDDs)	
Bandwidth	PCIe x2 or x4 lane	PCIe x2 or x4 lane	SATA 6 Gb/s
Form Factor	M.2/PCIe Expansion Card/U.2	M.2/PCIe Expansion Card	M.2/2.5" SSD
Max. Read Performance	>3000MB/s	>2000MB/s	>500MB/s
Max. Write Performance	>2000MB/s	>1500MB/s	>500MB/s

Table 1. Comparison of NVMe and SATA 6 Gb/s.
The following table compares bandwidths by generation. ATP's M.2 NVMe SSD is designed for a PCIe 3.0 interface and fits in a x4 lane, delivering up to 7.9 Gb/s transfer rate and up to 3.9 GB/s throughput.

PCIe				SATA		
Generation	Transfer Rate	Throughput per Lane		Generation	Transfer Rate	Throughput
Gen1	2.5 Gb/s	x1: 250 MB/s	x4: 1 GB/s	Gen1	1.5 Gb/s	150 MB/s
Gen2	4.9 Gb/s	x1: 500 MB/s	x4: 2 GB/s	Gen2	3 Gb/s	300 MB/s
Gen3	7.9 Gb/s	x1: 984.6 MB/s	x4: 3.9 GB/s	Gen3	6 Gb/s	600 MB/s
Gen4	15.8 Gb/s	x1: 1,969 MB/s	x4: 7.8 GB/s			

Table 2. Bandwidth Comparison of PCIe and SATA.

The following figure shows that while SATA delivers 4X the performance of mechanical HDDs, NVMe trumps SATA with 5X the sequential read and write performance, enabling the fastest performance and maximum theoretical speeds over any other storage protocol.

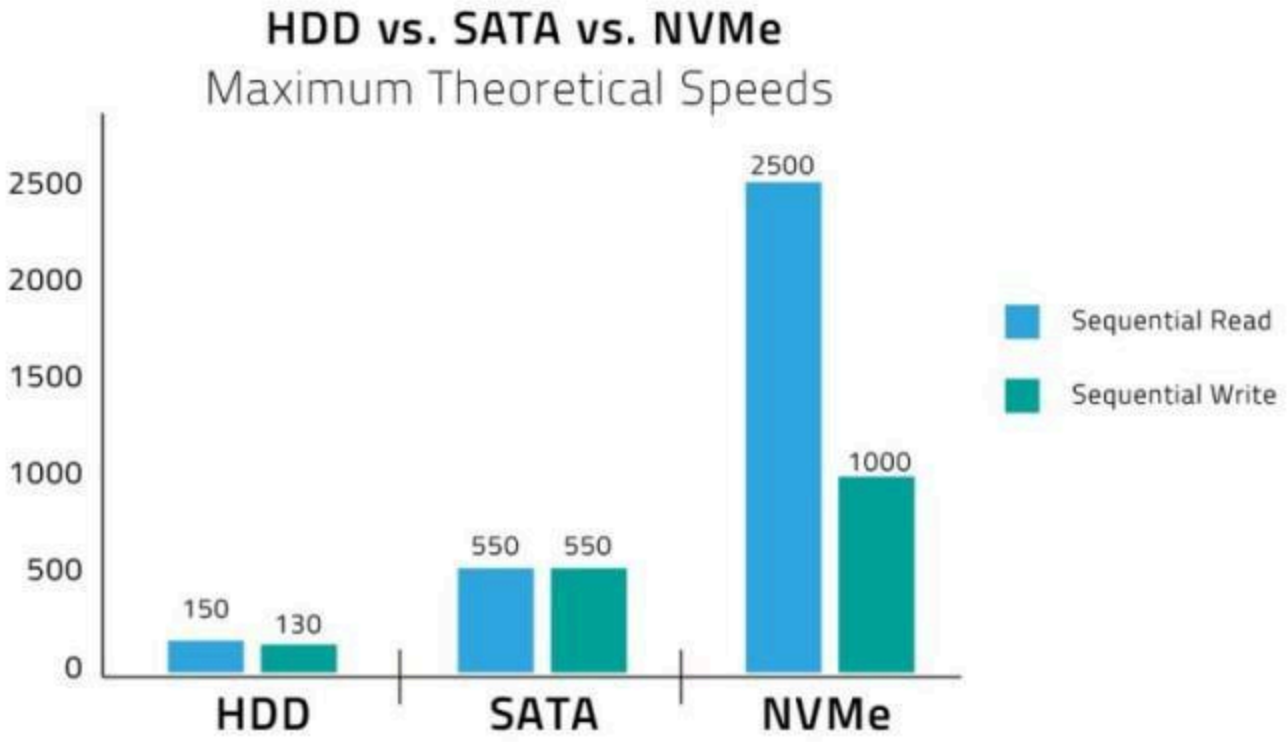


Figure 1. Maximum Theoretical Speed Comparison of HDD, SATA SSD and ATP NVMe SSD.

Pavilion compares RocE and TCP NVMe over Fabrics performance

TCP slower but not by much and enables Ethernet use

Chris MellorThu 16 Aug 2018 // 18:23 UTC

6_

Analysis Pavilion Data says NVMe over Fabrics using TCP adds less than 100µs latency to RDMA RoCE and is usable at data centre scale.

The biz is an NVMe-over-Fabrics (NVMe-oF) flash array pioneer and is already supporting simultaneous RoCE and TCP NVMe-oF transports.

Head of Products Jeff Sosa told *El Reg*: “We are …
supporting NVMe-over-TCP. The NVMe-over-TCP standard is ready to be ratified any time now, and is expected to be before the end of the year.

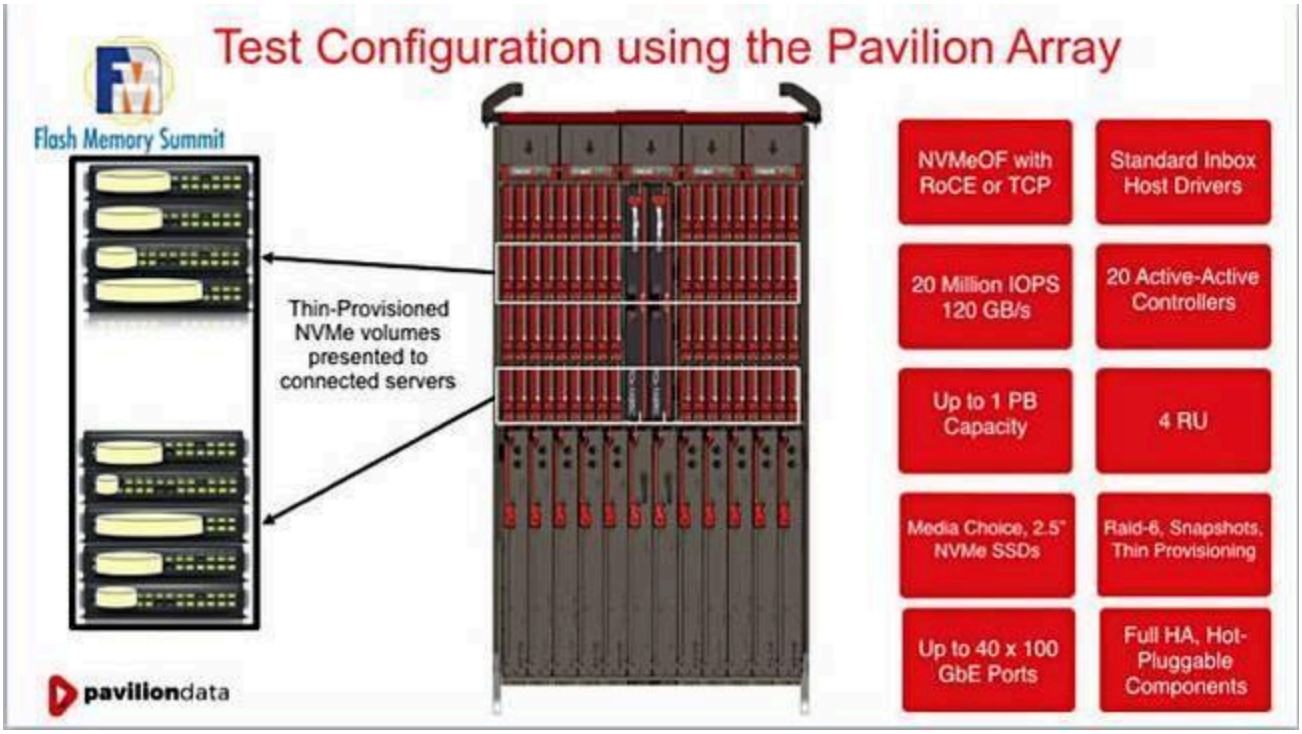
“We actually have a customer who is deploying both NVMe-oF with RoCE and TCP from one of our arrays simultaneously.”

Pavilion says NVMe-oF provides the performance of DAS, with the operational benefits of SAN. Its implementation has full HA and no single point of failure and says it offloads host processing with centralised data management.

That data management, when used for MongoDB for example, allows;

- Writeable clones to be instantly presented to secondary hosts, avoiding copying data over the network,
- Dynamically increase disk space size on-demand in any host,
- Instantly back up the entire cluster using high-speed snapshots,

- Rapidly deploy a copy of the entire cluster for Test/Dev/QA by using Clones,
- Eliminate the need for log forwarding by having each node write log data directly to a shared storage location,
- Orchestrate and automate all operations using Pavilion REST APIs.



Pavilion RoCE vs TCP NVMe-oF test setup. Click for larger image.

Pavilion compared NVMe-oF performance over RoCE and TCP with from one to 20 client accessors, and found average TCP latency was 183µs and RoCE's 107µs, TCP being 71 per cent slower.

A Pavilion customer NVMe-oF TCP deployment was data centre rather than rack scale with up to six switch-hops between clients and storage. It was focused on random write-latency serving 1,000s of 10GbitE non-RDMA (NVMe-oF over TCP) clients and a few dozen 25GbitE RDMA (NVME-oF with RoCE) clients.

The equipment included Mellanox QSA28 adapters enabling 10/25GbitE breakouts with optical fibre. There were 4xswitch ports consumed to connect 16xArray/Target ports physically cabled. Eight ports were dedicated to RDMA and 8 dedicated to NVMeTCP; both options equally "on the menu."

There were no speed-transitions between the storage array and 10GbitE or 25GbitE clients with a reduced risk of over-whelming port-buffers.

Early results put NVMeTCP (~200µs) at twice that of RoCEv2 (~100µs) but half that of NVMe-backed iSCSI (~400µs). On-going experimentation and tuning is pushing these numbers, including iSCSI, lower.

Pavilion produced a table indicating how RoCE and TCP NVMe-oF strengths differed;

	ROCE	TCP
Latency of Local SSDs	✓	
Standard / InBox Driver	✓	Coming Soon
No Hops or Direct-Connect	✓	✓
Inter-Rack / Aisle-Scale / DC-wide		✓
Streaming Workloads	✓	✓
Latency-Sensitive Workloads	✓	
CPU Intensive Application	✓	
Compatibility with Existing Servers & NICs		✓
NIC Bonding Required		✓

9

Suppliers supporting NVMe-oF using TCP as well as Pavilion include Lightbits, Solarflare and Toshiba (Kumoscale.) Will we see other NVMe-oF startups and mainstream storage array suppliers supporting TCP as an NVMe-oF transport? There are no signs yet but it would seem an easy enough (relatively) technology to adopt.

Pavilion's message here is basically, unless you need the absolute lowest possible access latency, then deploying NVMe-oF using standard Ethernet looks quite feasible and more affordable than alternative NVMe-oF transports - unless perhaps you run NVMe-oF over Fibre Channel.

BLOCK STORAGE VS. OBJECT STORAGE: WHEN TO USE EACH

WekaIO Inc.

There are multiple ways of storing data in computer systems. Some of the most common ways include using different types of file systems or block devices, using object stores, using different types of databases, and a host of other methods. Each approach has its advantages, disadvantages, and requirements, along with the use cases to which they will apply. Moreover, each approach has evolved in parallel to the technologies that enable it (e.g., block devices with Fibre channel, network filesystems with NFS protocol, etc.).

To compound the complexity, different storage media types exist, each with its own performance, durability, and price, such as Hard Disk Drives (HDDs), Solid State Drives (SSDs and NVMes), and more.
Storage companies have designed and built storage solutions implementing these approaches and adding on top of them, e.g. SAN (Storage Area Networks) appliances or All Flash NAS (Network Attached Storage) appliances, while IT and Storage admins would usually need to evaluate their internal workloads' needs and use the most suitable

storage appliance or even multiple different storage appliances for their different organization needs.

This blog takes a look at block storage and object storage, discusses the different technologies, and eventually discusses the relevant use cases of where each would apply.

What Is Block Storage?

As we know, computer data is written in units called “bits”: 8 bits are called a “byte,” 1024 bytes are called a “kilobyte” (KB), increasing in size with megabytes (MB), gigabytes (GB), terabytes (TB), petabytes)PB), and so on.
A “block” is actually several “bytes” of data grouped together. A traditional block was 512 bytes on older storage systems, while the currently more accepted block size is 4K.

For example, if you have a picture that is 128KB in size, that picture will be saved on 32 blocks of 4KB each. In this way any storage media (HDD, SDD, NVMe, other) can usually be exposed to the computer using it as a number of blocks (e.g., an 8TB HDD or SSD would be exposed as more than 2 billion blocks).

Since individual block devices such as HDDs and SSDs have only specific sizes and performance levels, storage companies created storage appliances that are composed of multiple block devices that can then be “exposed” to servers as a different number of block devices. These are called LUNs (Logical Unit Numbers).

For example, let’s consider a storage appliance with 10 X 10TB HDDs that is exposing a 1 X 100TB block device (LUN) to a server or 100 X 1.0TB block devices (LUNs) to 100 different servers. This scenario allows servers to “see” and use block devices that are bigger than any

single individual block device on the market. It also allows for better performance, as utilizing multiple devices concurrently provides higher performance than a single device, as well as additional sophisticated features (such as protection and more) that the storage appliance can implement.

A block storage appliance can expose multiple “block devices” (LUNs) to a server, while the server would work with these LUNs as if they were local only to it (that is, as if they are its local HDD or SSD). Therefore, these LUNs cannot be (easily) shared between different servers, and each server would see only its own assigned LUNs and no others.

Because a LUN is actually a group of blocks, the servers will usually (but not always) choose to create a file system on top of it, which means that some of the blocks in the LUN will be used for data and some for metadata (data about the data).

An example of that would be saving a 128KB image under the user’s pictures directory, using 128KB for the image and an additional 4KB for data about the image file (i.e., creation time, access time, permissions, etc.), as well as 4KB for the directory structure. An important distinction is that the block device is not aware of which of its used blocks include data and which include metadata; therefore, accessing this data can be used only through a server that is familiar with the file system that is used on this block device. That is, a LUN that is formatted with an ext4 filesystem can be used only by servers that are familiar with the ext4 filesystem, and even then, as mentioned above, no two servers will use this LUN concurrently.
Additionally, servers would need to be connected to the block storage appliances in some way. Traditionally that would be using Fibre Channel optical cables or using Ethernet connectivity.
Different protocols can be used here, mostly FCP or ISCSI, while more

modern NVMe-based block storage appliances will use NVMe over Fabrics (NVMe-oF).

Block Storage Advantages and Disadvantages

The ability to aggregate multiple block devices and expose them to multiple servers with different sizes and additional capabilities is a huge benefit to organizations that want to centralize all of their storage devices into a smaller number of block storage appliances connected using high-speed interconnects, yet there are some disadvantages as well.

Here are the benefits of block storage:

- **Centralization**: This case presents no need for multiple/different storage devices for servers, as each different storage device presents a challenge in terms of operations, cost, and protection. By centralizing all data into a smaller number of block storage appliances the IT and storage admins gain better control and utilization of the storage.
- **Backups and replication**: Because the data is centralized it is easy to use the block storage appliances' advanced features to create point-in-time snapshots of specific LUNs, to replicate LUNs to remote storage appliances, or even to use backup protocols such as NDMP to more efficiently back up the data.
- **Performance**: Because block storage appliances contain multiple devices and lay out the blocks across multiple devices in parallel, the performance that the servers experience from their LUNs is much faster than that of a single block device. These devices are considered to be the most performant storage type in terms of throughput and latency, as they perform only simple reads and writes and what was considered a high-performance interconnect.
- **Advanced features**: Some applications can implement advanced block commands to improve specific repeated operations on a block storage appliance. For example, the VMware(R) vSphere Storage APIs Array Integration (VAAI) enables the ability to tell the

storage to perform a copy operation or to write zeros to multiple block ranges without the need to go through the host (client).

Here are the disadvantages of block storage:

- **Dedicated interconnect layer**: Traditionally, block storage appliances eventually required their own interconnect network, which was Fibre channel, and each server that connected to the storage would need a Fibre Channel optical Host Bus Adapter (HBA). This also meant that the customer would need to purchase Fibre Channel Switches and deploy Fibre Channel cables, essentially creating a Fibre Channel interconnect fabric environment to be used only for accessing the storage. It was a very expensive purchase and very difficult to manage (multiple FC zones, zonesets, WWNNs, and WWPNs zoning). Newer configurations now use Ethernet instead of Fibre Channel, which simplifies operations somewhat.
- **Complicated management**: Eventually, block storage devices map specific LUNs to specific servers. The IT and Storage admin needs to make sure that all servers can see only their dedicated LUNs, as at some point two servers might write to the same LUN, which could result in data corruption.
- **Non Shared**: As mentioned previously, even if two servers are connected to the same storage appliance they can not easily share their LUN. This is even more problematic if there is a need to share data between different operating systems (for example, for a MAC on Windows and Linux to see the same data).

What Is Object Storage?

Many modern use cases require a location for data placement without the relationship of folders and directory structures. That's where object storage comes in handy. Additionally, sometimes there is no need for low-latency performance access to the data, while there is usually a need for high capacity at a very reasonable price.

One example might be a backup environment in which a backup application takes multiple files from a highly performant, expensive

environment and backs them up to a more reasonably priced environment. Another example could involve multiple surveillance cameras that are periodically sending the last several minutes of a recording as one file to a centralized-capacity location.

Object storage was designed for these use cases and more.

Simply put, an object store is a location that can accept and provide objects. An object can be a file or multiple files aggregated to a single object. Each object has a unique object name, but there is no directory or folder relationship between the objects. For example, a 128KB image can exist as an object with object name “image1.” It does not reside under any directory for any user or group. The permission scheme for

accessing objects is simpler than on file systems: a user just needs to provide the correct credentials to access an object.

The object storage appliance can retain metadata (data about the data) for every object. For example, all images that have a car in them can have the metadata tag “car” applied to them so that the object storage can then return all of the images with the tag “car.”

Object stores contain logical entities called buckets. While each bucket contains objects, a bucket can also be mirrored, or erasure coded, across multiple object storage appliances and data centers, as well as have its own credentials.

Object storage is most suited for large objects (objects that are in the MBs size or larger) and can sometimes provide high throughput, but still at a higher latency than block devices.
The most commonly used access protocol for object stores is called S3 (originally created by Amazon). This protocol describes a number of connectionless commands for objects, including PUT, GET, LIST, DELETE, and more. Recently, there has been a growing acceptance for the S3 protocol where applications can natively use S3 as their storage (without the need of a file system).

Object Storage Advantages

High-capacity, inexpensive storage can be used in multiple ways, even if its performance is not as fast as that of a block storage appliance.

- **Scalability**: Object storage appliances are built to accommodate for multiple petabytes of capacity with billions of objects, partially due to the fact that they do not need to accommodate for many things that filesystems do need to handle. Additionally, object storage appliances are very cost effective, with a price per gigabyte that makes them ideal for capacity tiering with some types of data.

- **Protection**: Data in the object store is usually erasure coded, which means that even if the object store increases in capacity, the data is still protected and can be serviced across multiple hardware failures that might occur–making it a reliable form of storage.
- **DR capabilities**: Object stores are built to mirror across data centers as well. Because there is no high performance expectation for the data, the fact that some of it should traverse between the WAN and the data center is not a limitation. This means that object stores are ideal for data types that should be available in case of a data center failure, allowing for continued work on DR in the data center.
- **IOT device support**: Because S3 protocol is a connectionless protocol, it is widely used with IOT devices that cannot create and maintain a file system connection (NFS, SMB, etc.) in situations when devices simply need to open a connection to upload or download objects and then close the connection.

When To Use Block Storage

Block storage is highly efficient in the following scenarios.

- **Single-server, high-performance**: For data that does not need to be shared across multiple servers and that is accessed from a single server (such as some DBs, Virtual environments, etc.) block storage provides fast, high-throughput, low-latency access, such as in databases and hypervisors running connections to block storage appliances.
- **Sync-mirror data**: Since most storage appliances support synchronous block mirroring between multiple storage arrays in different data centers, a highly used scenario for block storage is to have servers accessing data on a LUN in one data center that is block-sync mirrored to a remote data center. While on the remote data center there are servers connected to the replicated LUN, in case of a data center failure the remote LUN will become active, and applications will continue working on a remote data center, thereby ensuring business continuity even during a datacenter failure.

- **Centralization**: Sometimes it's essential to centralize data from multiple applications into a single location that is protected, resilient, and uses advanced features such as snapshots, deduplication, compression, and more. In many cases, even applications that are built to distribute across multiple local storage appliances (such as Cassandra, splunk, Elastic search) are being centralized on centralized storage for ease of management.

When To Use Object Storage

Object storage is ideal for the following workloads.

- **High scalability at cost effective price**: Object stores provide massive scalability with affordable economics.
  Backup and DR: Due to their efficient mirroring and erasure coding capabilities across data centers, as well as their versioning capabilities, object stores are suitable for backup and DR scenarios that require data to be restored or used following objects' deletion or data center failures.
- **Core-to-edge use cases**: Due to the connectionless properties of the S3 protocol, a common use case involves multiple-edge IOT environments that are pushing (and sometimes pulling) data from the core to a centralized object store. That data is then analyzed in the core environment, possibly analyzed directly on the object store or possibly copied first to a high-performance storage system.

The Future of Storage

An ideal storage appliance combines all of the above properties and more. It has the ability to provide a shared file system between multiple servers in a highly performant manner (similar to a block device but without all of the complication associated with managing multiple LUNs, mapping, and FC). Additionally, it has the economics and scalability of object storage–but without the performance limitations. WekaFS was designed with all of this in mind, with record breaking performance and scalability across NVMe devices, as well as scalability across object

storage appliances–all under a single namespace for capacity as well as backup and DR capability–while maintaining ease of use and management.

WHAT IS NETWORK FILE SYSTEM (NFS)?

Lynn Orlando. April 15, 2021

We are taught early in our lives that sharing is good. Network File System (NFS) was built on the principle of sharing. NFS is an Internet Standard, client/server protocol developed in 1984 by Sun Microsystems to support shared, originally stateless, (file) data access to LAN-attached network storage. As such, NFS enables a client to view, store, and update files on a remote computer as if they were locally stored. On the back end, NFS client software translates POSIX file access commands issued by applications into NFS server requests that respond with metadata, data, and status. The main versions in deployment these days (client and server) are NFSv3, NFSv4, and NFSv4.1.

WHAT IS NETWORK FILE SYSTEM (NFS)

Network File System (NFS), was a protocol invented in the 80's to facilitate remote file sharing between servers. There are multiple versions of NFS. NFS v3 is the most common. NFS is easy to use and manage and requires a client in the Kernel that supports NFS mounting.

Benefits of Using NFS

Over the years, NFS has evolved to support more security, better file sharing (locking), and better (caching) performance. Moreover, it's a

relatively affordable and easy-to-use solution for network file sharing that uses existing internet protocol infrastructure.

At present, here are the benefits of the NFS service:

- Multiple clients can use the same files, which allows everyone on the network to use the same data, accessing it on remote hosts as if it were acceding local files.
- Computers share applications, which eliminates the needs for local disk space and reduces storage costs.
- All users can read the same files, so data can remain up-to-date, and it's consistent and reliable.
- Mounting the file system is transparent to all users.
- Support for heterogeneous environments allows you to run mixed technology from multiple vendors and use interoperable components.
- System admin overhead is reduced due to centralization of data.
- Fewer removable disks and drives laying around provides a reduction of security concerns—which is always good!

How Does Network File System Work?

Fundamentally, the NFS client-server protocol begins with a "mount" command, which specifies client and server software options or attributes. NFSv4 is a stateful protocol, with over 30 unique options that can be specified on the mount command ranging from read/write block size, and protocol used. Security protocols validate client access to data files as well as data security options, etc.

Some of the more interesting NFS protocol software options include caching options, shared file locking characteristics, and security support. File locking and caching interact together, and both must be properly specified for shared file access to work. If file (read or write) data only resides in a host cache and some other host tries to access the same file, the data it reads could be wrong, unless both (or rather) all clients of

the NFS storage server use the SAME locking options and caching options for the mounted file system.

File locking was designed to support shared file access. That is when a file is accessed by more than one application or (compute) thread. Shared file access could be occurring within a single host (with or without multi-core/multi-thread) or across different hosts accessing the same file over NFS.

Disadvantages of Network File System

There are many challenges with the current NFS Internet Standard that may or may not be addressed in the future; for example, some reviews of NFSv4 and NFSv4.1 suggest that these versions have limited bandwidth and scalability (improved with NFSv4.2) and that NFS slows down during heavy network traffic. Here are some others:

- **Security**—First and foremost is a security concern, given that NFS is based on RPCs which are inherently insecure and should only be used on a trusted network behind a firewall. Otherwise, NFS will be vulnerable to internet threats.
- **Protocol chattiness**—The NFS client-server protocol requires a lot of request activity to be set up to transfer data. The NFS protocol requires many small interactions or steps to read and write data, which equates to a ton of overhead for someone actively interacting with today's AI/ML/DL workloads that consume a tremendous number of small files.
- **File sharing is highly complex**—Configuring and setting up proper shared file access via file locking and caching is a daunting task at best. On the one hand, it adds a lot of the protocol overhead, leading to the chattiness mentioned above. On the other hand, it still leaves a lot to be desired, inasmuch any each host's mount command for the same file system can easily go awry.

- **Parallel file access**—NFS was designed as a way to sequentially access a shared network file, but these days applications are dealing with larger files and non-sequential or parallel file access is required. This was added to NFSv4, but not a lot of clients support it yet.
- **Block size limitations**—The current NFS protocol standard allows for a maximum of 1MB of data to be transferred during one read or write request. In 1984, 1MB was a lot of data, but that's no longer the case. There are classes of applications that should be transferring GBs not MBs of data.

There are other problems with NFS, but these are our top five. Yes, block size restrictions could easily be made larger, but then the timeouts would need to be adjusted and perhaps rethought. And yes, parallel file access is coming, but protocol chattiness and file sharing (locking-caching) problems listed above are much more difficult to solve.

NFS has worked well for over 35 years now. It's unclear whether NFS can be salvaged in today's small file world. Yet another version of NFS could be pushed through the standard's committee, but our view is that the chattiness problem is too endemic in the protocol definition to be eliminated entirely, AND NFS either needs to fully support shared files or not, doing both is a prescription for failure.

How to accelerate Network File System performance?

NFS offers limited performance and scalability for modern environments. For today's networks and capabilities, it's very limited. NFS offers capacities of up to 1.5 Gb/s while network cards can offer 100 Gb/s. NFS is also not efficient in managing metadata. Weka offers the simplicity of NFS with the performance and scalability of SAN and the ability to saturate 100 Gb/s pipes.

NAS VS. SAN VS. DAS – ADVANTAGES & DISADVANTAGES

Lynn Orlando. July 15, 2020

Network Attached Storage

Network-attached Storage (NAS) is a file-level computer data storage server connected to a computer network. It provides storage and access to data from a central location to several authorized network users and other groups of clients. These systems are commonly used to support shared applications, including engineering software builds, data logging, email systems, video recording and editing, business analytics, financial records, genomics data sets and much more.

The file systems are contained in one or more storage drives often arranged into logical, redundant storage containers. The NAS sizes are dependent on speed, scale, and budget requirements. Therefore, they can be anything from a single desktop system all up to multi-petabyte scale-out systems. NAS uses one or more file access protocols that are exposed to an internal network. These are then presented by protocols such as NFS (network file system) or SMB (server message block), or a proprietary high-performance protocol that allow clients to attach to the NAS. Similar to TCP/IP devices and other computers, the NAS acts as a network node, maintains its IP addresses, and can effectively communicate with other network devices.

Advantages and Disadvantages of NAS

NAS systems are beneficial for small business owners because they are simple to operate; therefore, an IT professional is often not required. NAS is cost-effective with easy and secure data backup, and it can become the next step to DAS (direct-attached storage). It also significantly reduces wasted space over other storage technologies such as DAS or SAN (storage area network). Furthermore, NAS systems are continually accessible, making it easy for employees to collaborate, respond to customers, and support joint development projects. The NAS system also acts like the Cloud, where it can be accessed remotely using a network connection. Therefore, the employees can work from anywhere at any given time.

The weaknesses of NAS are related to scale and performance. The NAS is limited to its resources, and if the number of users requiring access increases, the NAS appliance cannot keep up, leading to slow performance and user frustration. The NAS systems cannot be easily scaled up or out, and NAS protocols such as Network File System (NFS) and SMB are not fast enough for high-performance applications due to the burdens of low throughput and high latency. Furthermore, NAS is network dependent, as files are shared over the local area network (LAN). The LAN transfers data from one place to another via data packets by dividing them into several segments and sending them to any terminals. However, any of those data packets can be delayed or be sent out of order, and traffic over LAN also becomes a determining factor.

Storage Area Network

A storage area network (SAN) provides block-level access for hosts that need control over their file systems. Prior to the development of SANs, a server would use the internal disk as a block device leveraging a local file system, but storage could only scale inside the server leading to islands of storage. A SAN sidesteps this issue and provides the block device across a network. Unlike NAS, SAN is a network storage system that requires complex technologies to maintain performance to the servers. The components can include gateway devices, dedicated switches (or VLANs on a shared network), tape backup units, controller nodes, and disk shelves. The SAN system uses protocols such as SCSI, iSCSI, and Fibre Channel.

Advantages and Disadvantages of SAN

Since SAN works on a separate network, it works similar to the direct-attached storage. That is, it can move the resources of the local area network, thereby creating a high-speed and organized environment that can be accessed by the operating system of each client. Therefore, it also allows for data storage quickly. Using a SAN means that the devices connected to the network need not use any local storage, which enables them to scale. If there is a need to move the SAN to another location, the data can be replicated rapidly, thus reducing the time for the recovery process.

Since a SAN is made up of elaborate and sophisticated interconnected devices, it has its disadvantages due to the implementation complexities. In addition, while the storage pool is shared among many servers, each is highly complex and can become an expensive investment.

Direct Attached Storage

Direct attached Storage (DAS) is digital storage that is directly connected to the system (i.e., a PC or a server) through an internal cable. The DAS system holds multiple hard disk drives in a single enclosure, which is directly connected to a machine through an HBA (Host Bus Adapter). Between these disk drives there is no network device (i.e., a switch, a hub, or a router, or network cable). For an individual PC user, the system's hard disk is the standard form of DAS. However, in enterprises or businesses, the separate disk drives in one server and the drives external to that server are either attached directly or attached through Small Computer System Interface (SCSI), Serial Advanced Technology Attachment (SATA), or Serial Attached SCSI (SAS). Another limitation of DAS is that data cannot be shared with different servers or users.

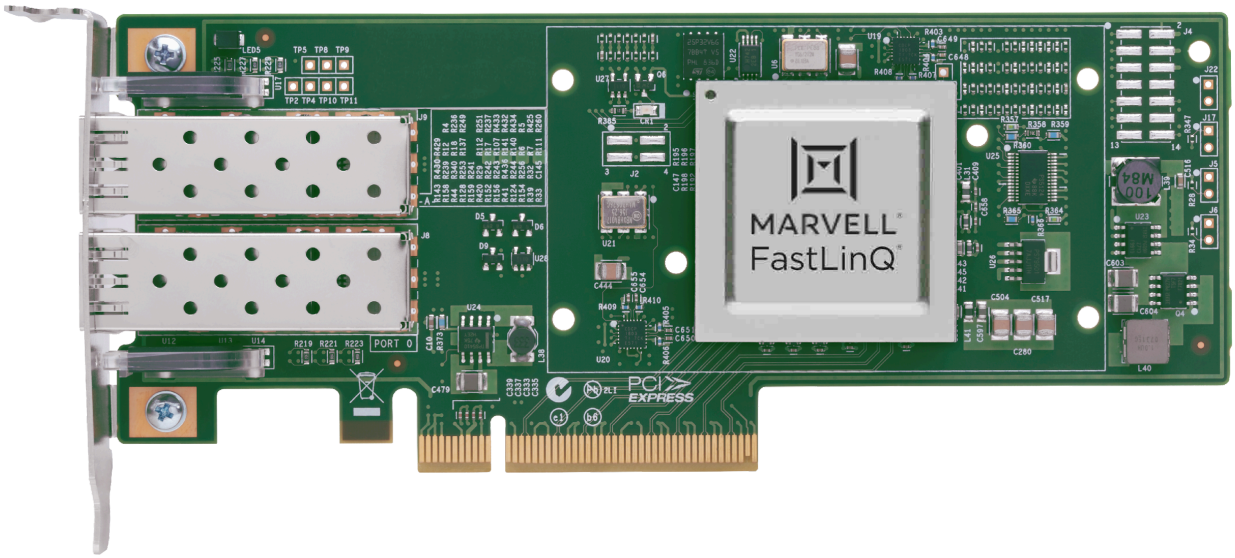
Advantages and Disadvantages of DAS

DAS systems are simple to use, and the technology is widely available. It provides great performance compared to SAN or NAS, but the storage capacity cannot be expanded. The disadvantages are that different user groups cannot access the data; it is only directly accessible from the applications running in the individual server or desktop machine. Furthermore, DAS does not incorporate any network hardware nor a related operating environment to provide a facility to share storage resources independently.

Small enterprises favor NAS systems as active, scalable, and economical storage solutions. SANs are high-performing multifaceted systems, and they are ideal for companies looking for high efficiency and dependability. DAS systems are used by individual small organizations or often to improve performance in an enterprise by copying data from

the NAS system to the local drives inside the DAS machine. While this is more performant, it can be a headache to manage the logistics of local copy.

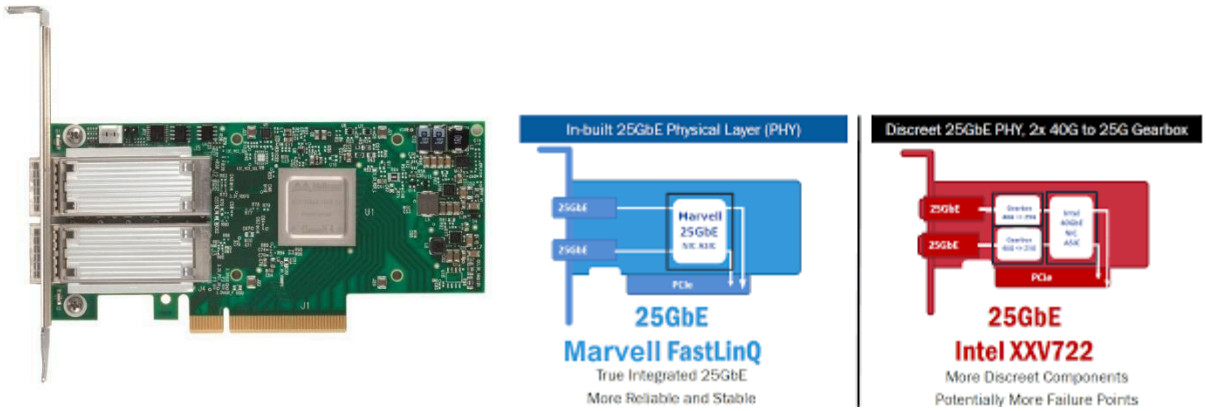
CNA storage NICs

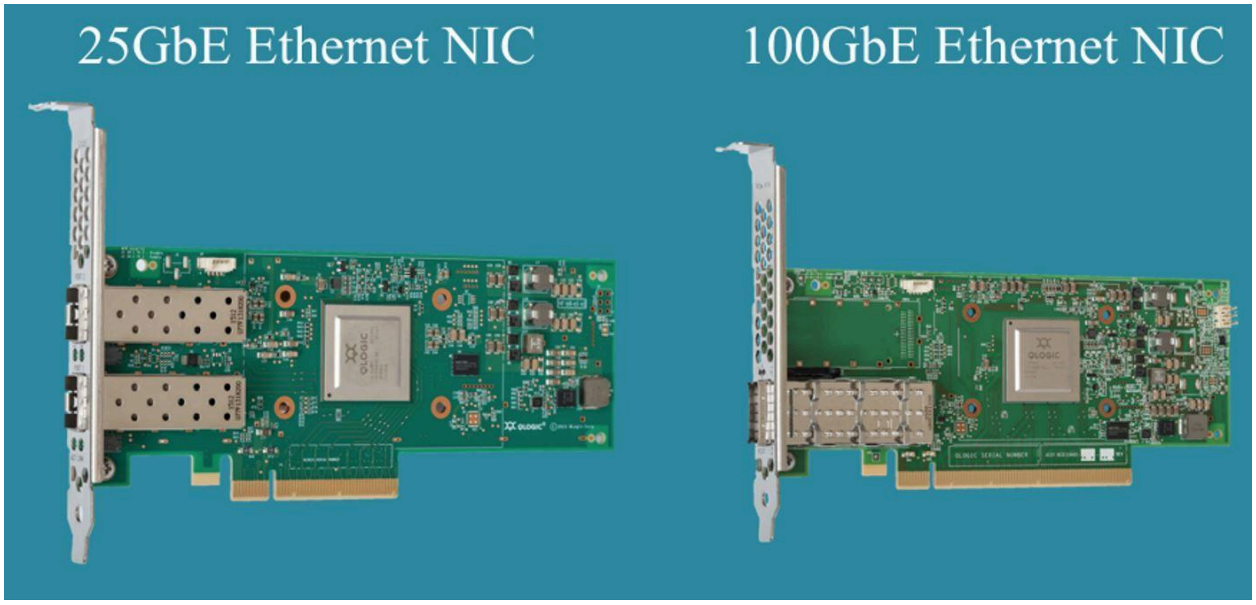


The Marvell® FastLinQ®41000/45000/8400 Series converged network adapters (CNAs) comprise multiple generations of high-performance CNAs ideally suited for enterprise-class data centers, public and private clouds, managed service providers (MSPs) and telco deployments. This FastLinQ Series supports 10/25/40/50/100Gb Ethernet (10/25/40/50/100GbE) and includes key features like Universal RDMA (concurrent RoCE and iWARP), server virtualization with NIC partitioning (NPAR) and single root IO virtualization (SR-IOV), network tunneling with stateless offload for Virtual Extensible LAN (VXLAN), Network Virtualization using Generic Routing Encapsulation (NVGRE), Generic Routing Encapsulation (GRE) and Generic Network Virtulization Encapsulation (GENEVE), and network storage with iSCSI and Fibre Channel over Ethernet (FCoE) offloads.

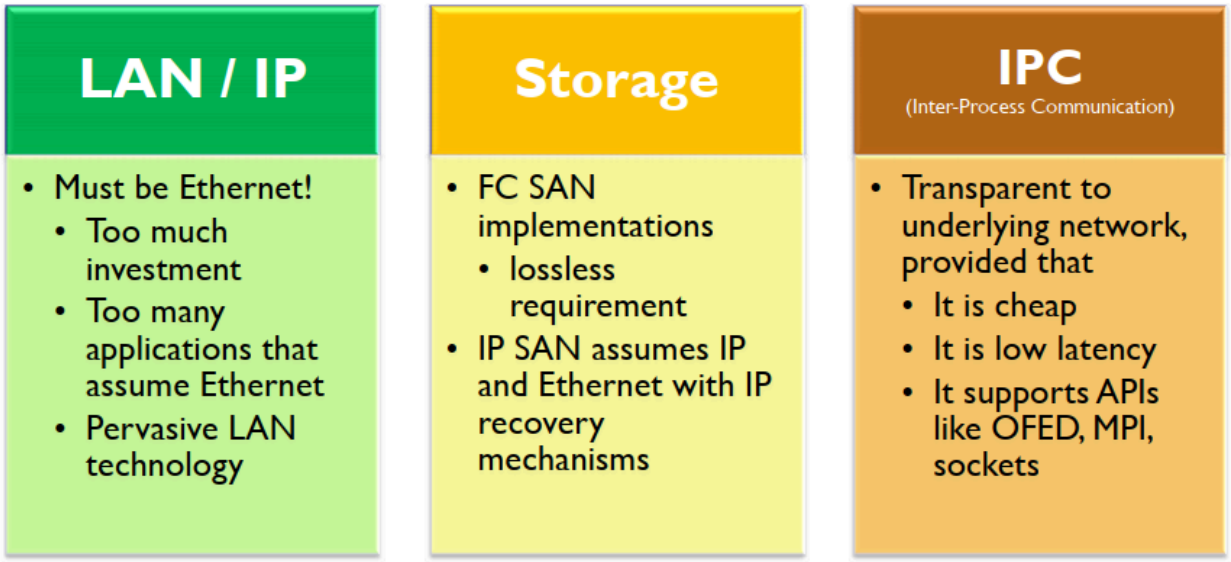
Key Features

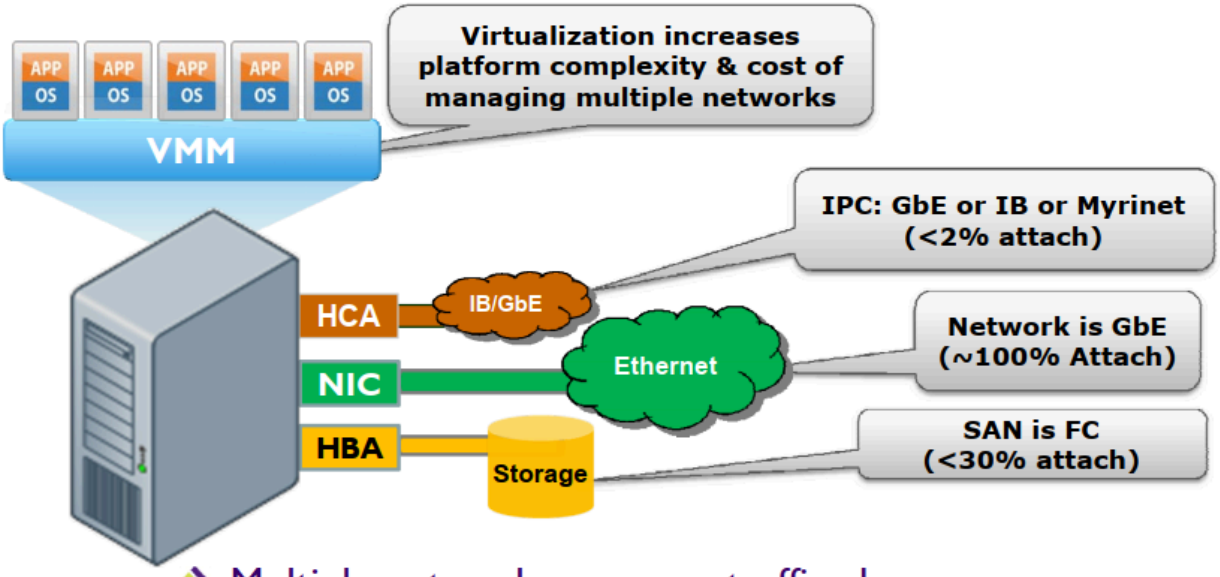
- Industry-leading networking performance
- Full hardware offload for FCoE and iSCSI protocol processing
- Support for multiple concurrent protocols: L2, RDMA (41000/45000 Series Only), FCoE, iSCSI and Fibre Channel over Ethernet (FCoE) offloads
- VLAN, teaming, jumbo frame, and stateless offload support
- NIC Partitioning (NPAR) that works with any Ethernet switch
- Managed by QConvergeConsole and native OS-based management tools
- Bullet-proof FCoE and iSCSI drivers that are common across all adapters
- Compatible with existing Fibre Channel and iSCSI storage
- PXE, UEFI, and iBFT boot support





- **Data Center Bridging (DCB)** is an architectural collection of Ethernet extensions designed to improve Ethernet networking and management in the Data Center.
- Sometimes also called
  - CEE = Converged Enhanced Ethernet
  - DCE = Data Center Ethernet (Cisco Trademark)
  - EEDC = Enhanced Ethernet for Data Centers



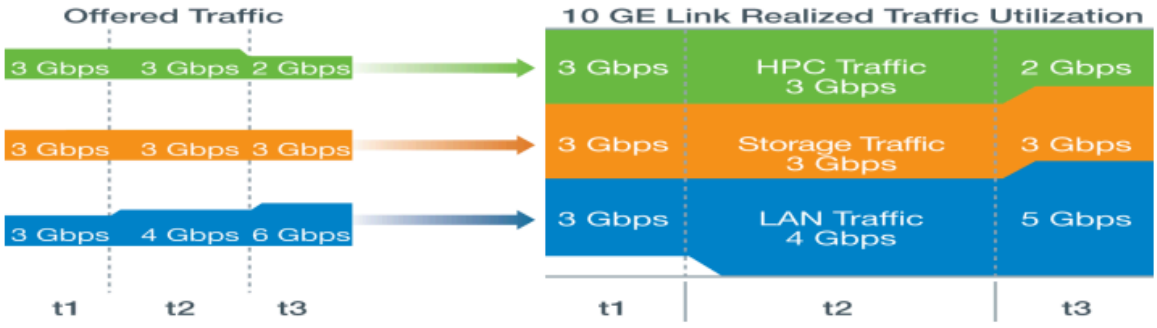


- Multiple networks, one per traffic class
  - IPC over an InfiniBand network
  - IP and other LAN protocols over an Ethernet network
  - SAN over a Fibre Channel network

Ethernet Enhancements for Storage

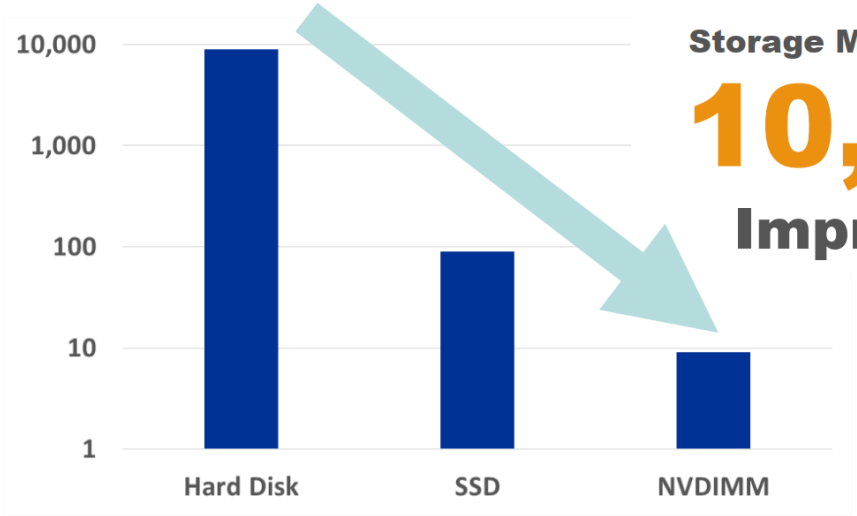
6

Priority based Bandwidth Management



Enables Intelligent sharing of bandwidth between traffic classes control of bandwidth

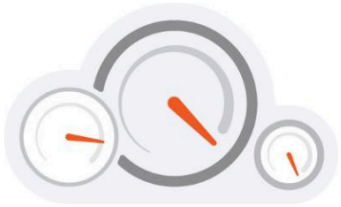
Storage Media Technology



Storage Media Access Time

10,000X

Improvement



SDC 17

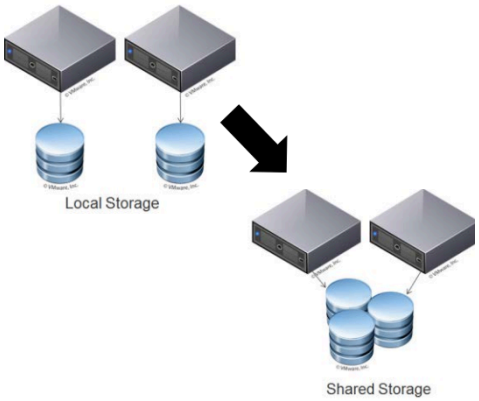
2

Ethernet Storage Fabric

PERFORMANCE	INTELLIGENCE	EFFICIENCY
• Highest Bandwidth • Lowest latency • RDMA and storage offloads • Native NVMe-oF Acceleration	• Integrated & Automated Provisioning • Hardware-enforced Security & Isolation • Monitoring, Management, & Visualization • Storage-aware QoS	• Just Works Out of the Box • Flexibility: Block, File, Object, HCI • Converged: Storage, VM, Containers • Affordable: SAN without the $$

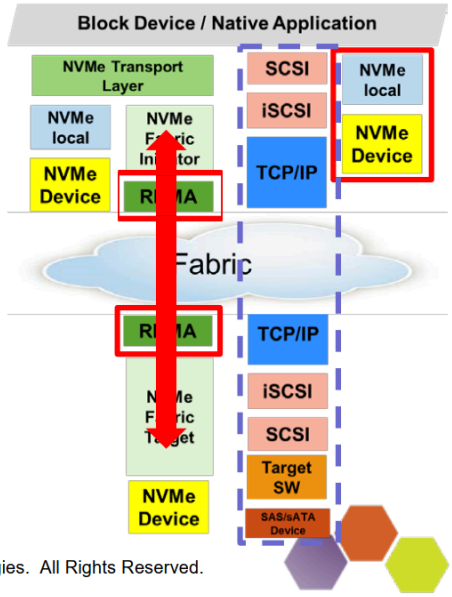
Extending NVMe Over Fabrics (NVMe-oF)

- NVMe SSDs shared by multiple servers
  - Better utilization, capacity, rack space, power
  - Scalability, management, fault isolation
- NVMe-oF industry standard
  - Version 1.0 completed in June 2016
- RDMA protocol is part of the standard
  - NVMe-oF version 1.0 includes a Transport binding specification for RDMA
  - Ethernet(RoCE) and InfiniBand

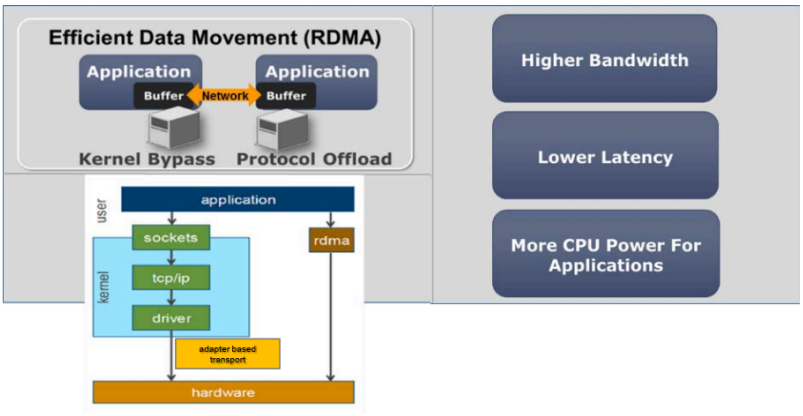


How Does NVMe-oF Maintain Performance?

- Extends NVMe efficiency over a fabric
  - NVMe commands and data structures transferred end to end
- RDMA is key to performance
  - Reduces latency
  - Increased throughput
  - Eliminates TCP/IP overhead



RDMA: More Efficient Networking

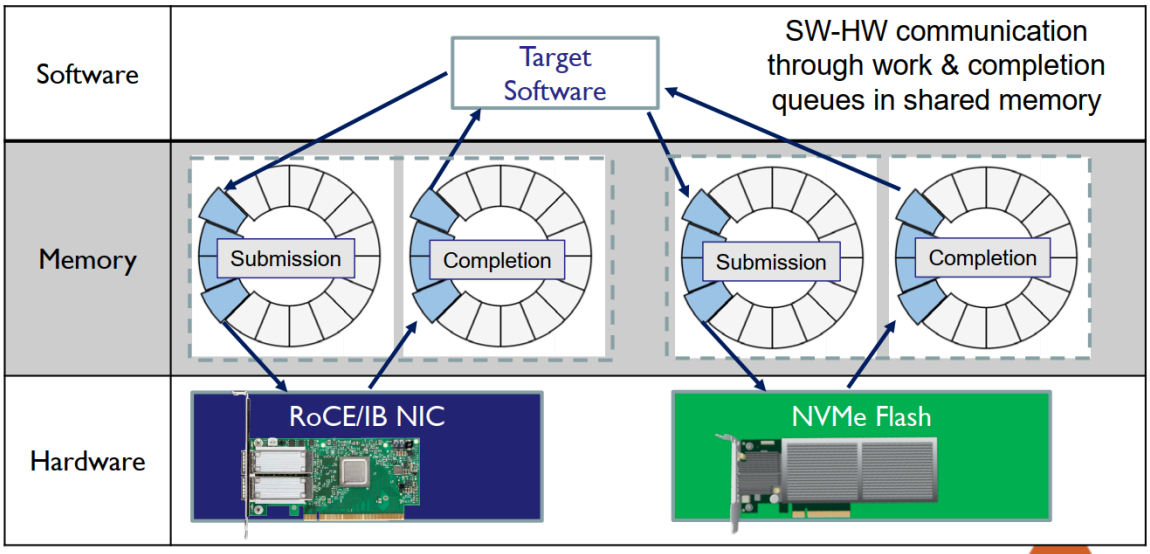


RDMA Performs Four Critical Functions in Hardware

1. Reliable Data Transport
2. App-level user space I/O - AKA: Kernel Bypass
3. Address Translation
4. Memory Protection

- CPU not consumed moving data - Free to run apps!

RDMA is Natural Extension for NVMe



Server Disaggregation: Sometimes the Sum of the Parts Is Greater Than the Whole

By Bill Dawkins | November 20, 2017
https://www.delltechnologies.com/en-us/blog/server-disaggregation-sometimes-the-sum-of-the-parts-is-greater-than-the-whole/

The notion of "the whole being greater than the sum of its parts" is true for many implementations of technology. Take, for example, hyper-converged infrastructure (HCI) solutions like the Dell EMC VxRail. HCI combines virtualization software and software defined storage with industry standard servers. It ties these components together with orchestration and infrastructure management software to deliver a combined solution that provides operational and deployment efficiencies that, for many classes of users, would not be possible if the components were delivered separately.

However, certain challenges require separating out the parts – that's where the solution is found. And, that is true in the case of Server Disaggregation and the potential benefits such an architecture can provide.

So, what is Server Disaggregation? It's the idea that for data centers of a certain size, efficiencies of servers can be improved by dissecting the traditional servers' components and grouping like components into resource pools. Once pooled, a physical server can be aggregated (i.e., built) by drawing resources on the fly, optimally sized for the application it will run. The benefits of this model are best described by examining a little history.

B.V.E. *(Before the Virtualization Era)*

Before virtualization became prevalent, enterprise applications were typically assigned to physical servers in a one-to-one mapping. To prevent unexpected interactions between the programs, such as one misbehaving program consuming all the bandwidth of a server component and starving the other programs, it was common to give critical enterprise applications their own dedicated server hardware.

Figure 1 describes this model. Figure 1 (a) illustrates a concept physical server with its resources separated by class type: CPU, SCM[1], GPU and FPGA, Network, Storage. Figure 1 (b) shows a hypothetical application deployed on the server and shows the portion of the resources the application consumed. Figure 1 (c) calls out the portion of the server's resources that were underutilized by the application.

Figure 1 (c) highlights the problem with this model, overprovisioning. The underutilized resources were the result of overprovisioning of the server hardware for the application to be run. Servers were overprovisioned for a variety of reasons including lack of knowledge of the application's resource needs, fear of possible dynamic changes in workload, and to account for anticipated application or dataset growth overtime. Overprovisioning was the result of a "better safe than sorry" mindset, which was not necessarily bad philosophy when dealing with mission critical enterprise applications. However, this model had its costs (e.g., higher acquisition costs, greater power consumption, etc.). Also, because the sizing of multiple servers for applications was done when the servers were acquired, a certain amount of configuration agility was removed as more knowledge about the true resource needs of the applications was learned. Before virtualization, data center server utilizations could be as low as 15% or less.

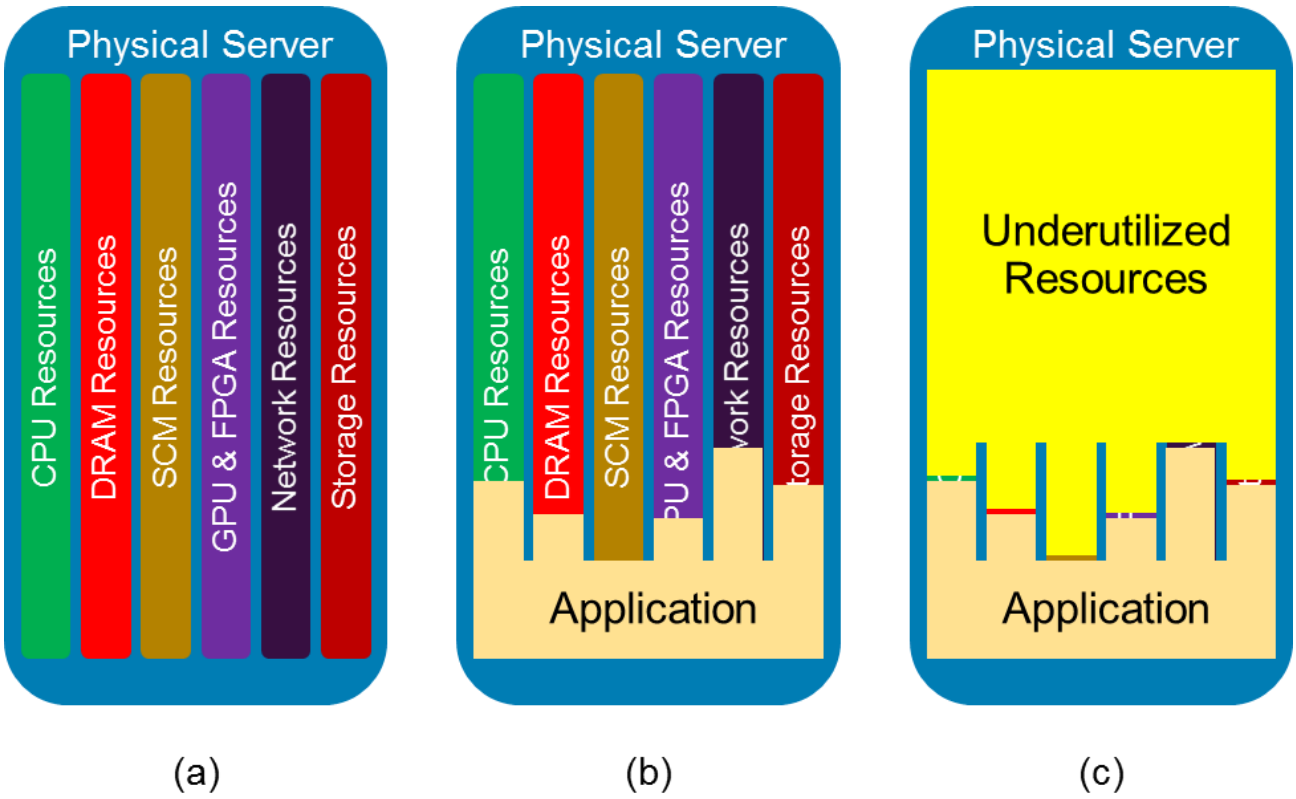


Figure 1: Enterprise Application Deployment before Virtualization

The Virtualization Age

When virtualization first started to appear in data centers, one of its biggest value propositions was to increase server utilizations. (Although, many people would say, and I would agree, that equally important are the operational features that virtualization

environments like VMware vSphere provide. Features like live-migration, snapshots and rapid deployment of applications, to name a few.) Figure 2 shows how hypervisors increased server utilizations by allowing multiple enterprise applications to share the same physical server hardware. After virtualization was introduced to the data center server utilizations could climb to 50% to 70%.

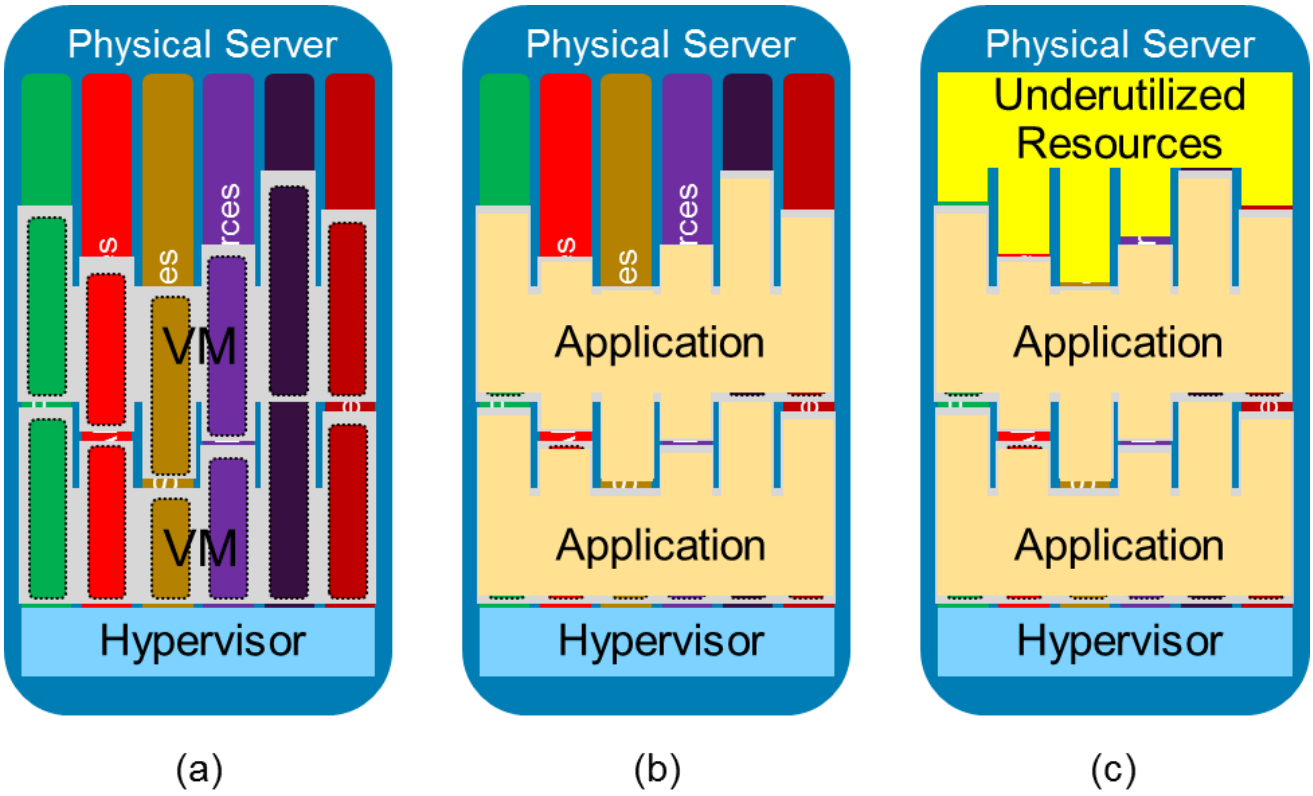


Figure 2: Enterprise Application Deployment after Virtualization

Disaggregation: A Server Evolution under Development

While the improvement of utilization brought by virtualization is impressive, the amount of unutilized or underutilized resources trapped on each server starts to add up quickly. In a virtual server farm, the data center could have the equivalent of one idle server for every one to three servers deployed.

The goals of Server Disaggregation are to further improve the utilization of data center server resources and to add to operational efficiency and agility. Figure 3 illustrates the Server Disaggregation concept. In the fully disaggregated server model, resources typically found in servers are grouped together into common resource pools. The pools are connected by one or more high-speed, high-bandwidth, low latency fabrics. A

software entity, called the Server Builder in this example, is responsible for managing the pooled resources and rack scale fabric.

When an administrator or a higher-level orchestration engine needs a server for a specific application, it sends a request to the Server Builder with the characteristics of the needed server (e.g., CPU, DRAM, persistent memory (SCM), network, and storage requirements). The Server Builder draws the necessary resources from the resource pools and configures the rack scale fabric to connect the resources together. The result is a disaggregated server as shown in Figure 3 (a), a full bare-metal, bootable server ready for the installation of an operating system, hypervisor and/or application.

The process can be repeated if the required unassigned resources remain in the pools, allowing new servers to be created and customized to the application to be installed. From the OS, hypervisor or application point of view, the disaggregated server is undistinguishable from a traditional server, although with several added benefits that will be described in the next section. In this sense, disaggregation is an evolution of server architecture, not a revolution as it does not require a refactoring of the existing software ecosystem.

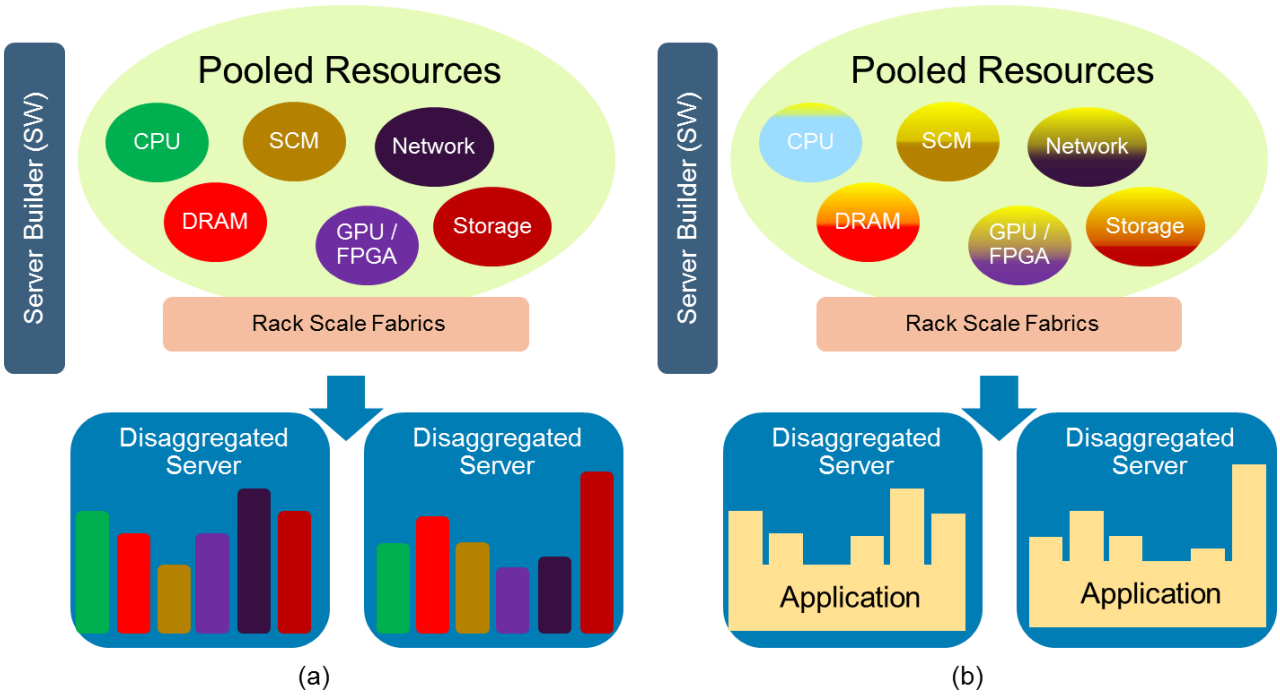


Figure 3: Disaggregated Servers

The Benefits of Being Apart

While having all the capabilities of a traditional server, the disaggregated server has many benefits:

- *Configuration Optimization:* The Server Builder can deliver a disaggregated server specifically composed of the resources a given application requires.
- *Liberation of Unused Resources*: Unused resources are no longer trapped within the traditional server chassis. These resources are now available to all disaggregated servers for capability expansion or to be used for the creation of additional servers (see Figure 3 (b)).
- *Less Need to Overprovision:* Because resources can be dynamically and independently added to a disaggregated server, there will be less temptation to use a larger than needed server during initial deployment. Also, since unused resources are available to all existing and future configurations, spare capacity can be managed from a data center level instead of a per server level, enabling a smaller amount of reserved resources to provide the overflow capacity to more servers.
- *Independent Acquisition of Resources:* Resources can be purchased independently and added separately to their respective pools.
- *Increased RAS (Reliability, Availability and Serviceability):* High-availability can be added to server resources where it was not possible or economical to do so before. For example, the rack scale fabric can be designed to add redundant paths to resources. Also, when a CPU resource fails, the other resources can be remapped to a new CPU resource and the disaggregated server rebooted.
- *Increased Agility through Repurposing:* When a disaggregated server is retired, its resources return to the pool which in turn can be reused in new disaggregated servers. Also, as application loads change, disaggregated servers devoted to one application cluster can be reformed and dedicated to another application cluster with different resource requirements

The above list is not exhaustive and many other benefits of this architecture exist.

The Challenges (and Opportunities) of a Long(ish)-Distant Relationship

Full server disaggregation is not here yet and the concept is under development. For it to be possible, an extremely low-latency fabric is required to allow the components to be separated at the rack level. The fabric also needs to support memory semantics to be able to disaggregate SCM (Storage Class Memory). It remains to be seen if all DRAM can be disaggregated from the CPU, but I believe that large portions can depending on the requirements of the different classes of data used by an application. Fortunately, the industry is already developing an open standard for a fabric which is perfect for full

disaggregation, Gen-Z. Information about the Gen-Z effort can be found at www.genzconsortium.org.

The software that controls resources and configures disaggregated servers, the Server Builder, needs to be developed. It also provides opportunities for the addition of monitoring and metric collection that can be used to dynamically manage resources in ways that were not possible with the traditional server model.

Another opportunity is the tying together of the disaggregated server infrastructure with the existing orchestration ecosystems. Server Disaggregation is in no way a competitor to existing orchestration architectures like virtualization. On the contrary, Server Disaggregation is enhancing the traditional server architecture that these orchestration environments already use.

One can imagine that the management utilities administrators use to control their orchestration environments could be augmented to communicate directly to the Server Builder to create the servers they need. The administrator may not ever need to interface directly to the Server Builder. The benefits of disaggregation should be additive to the benefits of the orchestration environments.

Conclusion: An Exciting Time in Server Architecture

It is an exciting time to be involved in server architecture. New technologies like SCM and rack scale, low-latency fabrics are opening new doors for server innovation. Server Disaggregation has the potential to be one of these important innovations. Indeed, we have already seen some of the benefits of the disaggregation of some of the server components in systems like the Dell EMC PowerEdge FX2 and Dell EMC PowerEdge VRTX. Server Disaggregation can build on the benefits these examples provide and lead to a more efficient and more dynamic server infrastructure environment.

[1] SCM – Storage Class Memory. A class of emerging persistent memory technologies with latencies lower than NAND flash.

NVMe SSD Classification

NVM Express® technology is driving the next generation of SSDs for data centers. NVMe® SSDs are not limited by legacy form factors or protocols, and they can better address the needs of servers and storage in enterprise and hyperscale data centers. Each of these use cases has distinct requirements and application environments, and they are outlined in the table below. There will always be exceptions to these guidelines, but these are generally fit profiles for NVMe SSD classification.

More details on each of the classification use cases are provided in a new NVMe SSD Classification White Paper which describes how NVMe SSDs are used in enterprise servers, enterprise storage, data center/hyperscale servers, and data center/hyperscale storage environments. This classification may also apply to SATA and SAS SSDs.

NVMe SSD Classification	Enterprise Server	Enterprise Storage	Data Center / Hyperscale Server	Data Center / Hyperscale Storage
Applications / Use Cases	Traditional DAS/RAID CRM/ERM/collaborative/ HCI	AFA Cache	Boot Cloud data storage Server data storage	Durable storage RAID EC JBOD/JBOF

Form Factors	M.2->E1.S 2.5-inch (U.2) E3.S, E3.L AIC	M.2->E1.S 2.5-inch (U.2) E3.S E3.L	M.2->E1.S U.2 (7 &15 mm-Z)	U.2 E1.L E1.S
Power	8.25-11W (M.2) 12-25W (2.5-inch/U.2) 25-40W (E1.S, E1.L)	8.25-11W (M.2) 12-25W (2.5-inch/U.2) 25-40W (E1.S, E1.L)	8.25-11W (M.2) 12-25W (E1.S) 12-25W (U.2)	12-25W (E1.S) 25W (E1.L)
Performance	high performance, read intensive and mixed workload class drives, and best in class performance	high performance, mixed workload, and best in class performance	IOPS/TB scaling, mixed workloads	read bandwidth, seq write (ZNS)
Latency	Care about ave latency. Better than SAS/SATA	Mixed workload QoS	Read QoS, tail latencies	read priority
Endurance-Drive Writes per Day (DWPD)	1 DWPD (Read-Intensive) 3 DWPD (Mixed Use)	1 DWPD (Read-Intensive) 3 DWPD (Mixed Use)	<=1 DWPD	<1 DWPD

	5-10 DWPD (Write-Intensive)			
Capacity	1-8TB (2021)	2-30TB (2021)	1-2TB (2020) > 4TB 2021-22	4-8TB TLC, >8TB QLC
Management	NVMe-MI (OOB) Swordfish (NVMe-oF)	NVMe-MI (OOB) Swordfish (NVMe-oF)	NVMe CLI (IB via SW) NVMe-MI basic for temp only (BMC)	NVMe CLI (IB via SW)
Features	Security Manageability Single-port, RAID	Dual-port (2.5-inch form factor) X8/16 PCIe AIC Higher prices (greater than SANs) Capacities up to 30TB	Single-port QoS Sets / IOD Latency TRIM performance Multi-tenancy	ZNS SDS
SSD Vendor Examples	Enterprise Server	Enterprise Storage	Data Center / Hyperscale Server	Data Center / Hyperscale Storage
KIOXIA	CM6	CM6	XD6, CD6, CD7	XD6, CD6, CD7
Intel	P4510 (3.0), P5500 (4.0)	D4510 (3.0)	P4510 (3.0)	P4326 (QLC, 3.0)
Micron	7300 PRO, MAX	7300 MAX	7300 PRO	9300 PRO

	9300 PRO, MAX			
Samsung	PM1733/5, SZ1735	PM1733/5, SZ1735	PM9A3	PM9A3

Use High Speed Direct Attach Cable for Data Center Interconnection

https://community.fs.com/blog/use-direct-attach-cable-assemblies-for-data-center-interconnection.html

Margaret

August 23, 2021

High speed direct attach cable (DAC) assemblies, or twinax cable assembly used in data centers. It provides a lower cost, higher density alternative cabling solution for high speed 10G-400G interconnect to fiber optic transceivers. To further comprehend their features, let us walk you through some insights in respect to what DAC cables are and power high speed data center interconnect.

What Is DAC Cable?

What is DAC cable? A direct attach cable, short for DAC cable, is a type of factory terminated twinax cable that connects directly into "transceivers" and can be brought in several lengths for short distances of up to 15 meters. These cables consist of fixed lengths of shielded copper coaxial or fiber optic cable with fixed pluggable transceivers on either end with varying gauges from 24 to 30AWG. DAC cable can achieve interconnection up to 15m via twinaxial copper cable, and reach 100m over active optical cable at 10Gbps, 40Gbps or beyond.

There are three common DAC cable Types:

- Passive DAC Cables
- Active DAC Cables
- Active Optical cables

Passive vs Active DAC Cables

High speed direct attach cable can be classified into direct attach copper cable(DACC) and direct active optical cable (DAOC). Read AOC cables for the detailed information.

DACC cables can be either passive or active. Passive DAC cable contains no electrical components, so it has minimal power consumption of <0.15w, but the linking distance is limited to 7m. While active DAC cable contains electrical components in the connectors that can boost signal levels, allowing reaching greater distances (5m or more) via copper media and ensure better transmission quality. This makes active copper cables slightly more expensive and consumes more power than direct attach passive copper cables.

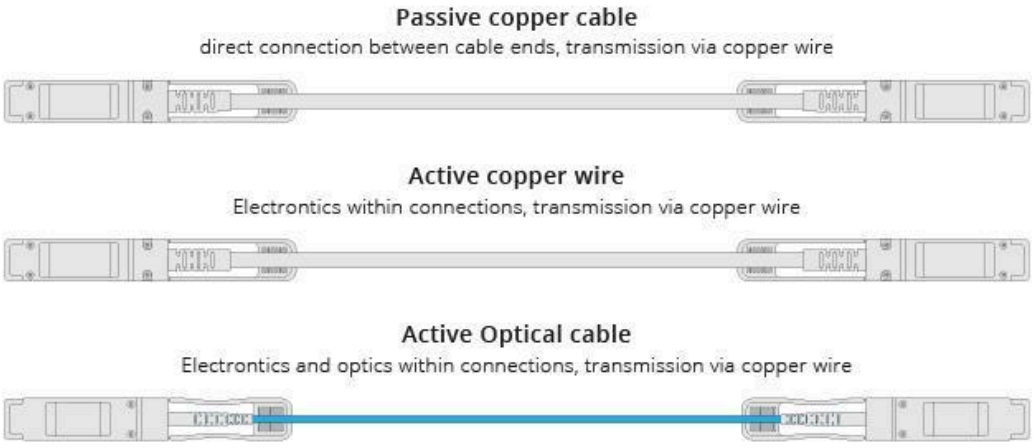


Figure 1: Passive or active DAC cables

Direct active optical cable bonds the fiber connection inside the transceiver ends, creating a complete fiber cable assembly much like a direct attach copper cable, but with a 15 meter reach capability, 2x improvement over a passive DACs limit. Also, the additional circuitry of the ACC does increase its power consumption to around 0.5-1.0 W, on average.

Direct Attach Cable vs. Optical Transceiver Module

DAC, AOC cables and optical transceiver modules are used to connected switches with one another when creating a stack or switches to routers or servers. Optical transceiver modules need fiber cables to be plugged into the transceiver module, while DAC and AOC cables are the fixed cable assemblies with different form factor connectors on both ends. Read DAC AOC cables comparison for detailed information.

	Reach	Cable Type	Power Consumption	Weight	Bend Radius	Application	Price

High Speed Direct Attach Cable	Around 1m-100m	Twinax copper cable or fiber optic cable	0.15w-1w	DAC:Heaviest AOC: Lightest	24 AWG-30 AWG	ToR, Adjacent racks, Middle of Row, End of Row, Zone to Zone	Low
Optical Transceiver Module	Per TIA/IEEE standard (Up to 160km)	Twinax copper cable or fiber optic cable	Up to 4.5w	Depend on transceivers & cabling used	Depend on cabling used	ToR, Middle of Row, End of Row, Zone to Zone	High

According to the table, it's obvious that high speed direct attach cable costs less, consume less power. But it usually has a limited distance compared to transceiver attached with cabling. Thus for any short range connection measuring from 5m to 100m, a better performing DAC cables offer easier and more affordable solution. But when the distance range is beyond 100m, optical transceivers are recommended.

Power High Speed Data Center Interconnect

High speed direct attach cables comes in many different configurations linking both new and older equipment, which optimize costs at every connection point. DAC cables are used in 32-to-56 port top-of-rack switches supporting up to 128 links (4x25G times 32 ports). Usually higher data rate DAC cable is backwards compatible to lower data rate DAC twinax cables. For example, 10G SFP+ DAC cables is backward compatible to 1 Gb/s.

DAC cables in different form factors are as follows:

- 10G SFP+ cables
- 25G SFP28 cables
- 40G QSFP+ cables
- 56G QSFP+ cables
- 100G QSFP28 DAC
- 200G QSFP56 DAC
- 400G QSFP-DD DAC
- QSFP-4SFP Breakout cables
- QSFP28-4SFP28 Breakout cables
- 200G DAC Breakout Cables
- 400G DAC Breakout Cables

Shop Passive or active dac cables at FS.com.

Figure 2: FS all series of Direct Attach Cables

What's new - 400G QSFP-DD DAC Cables

400G direct attach cables for short-distance DCI (Data Center Interconnect) have been mass-produced and put into market, which includes 400G DAC and 400G AOC. For more tech information about 400G DAC cables.

Data Center Cabling Solution: DAC Cables vs AOC Cables

Worton
October 28, 2021

https://community.fs.com/blog/guide-to-10g-dac-and-aoc-cables.html#:~:text=AOC%20Cables%20Comparison-,DAC%20cables%20are%20used%20to%20connect%20switches%2C%20servers%2C%20and%20storage,different%20racks%20inside%20data%20centers.

DAC Cables and AOC Cables are widely applied in data centers for high-performance computing network cabling system owing to their lower latency, lower power and lower cost. DAC Cables and AOC Cables come in a variety of configurations to meet network requirements. Each is available in 10G SFP+, 25G SFP28, 40G QSFP+, and 100G QSFP28 data rates with additional options for breakouts from 40G to 4x10G or 100G to 4x25G variants.

DAC/AOC Basics and Types

Direct Attach Cable (DAC) is comprised of a twinax copper cable terminated with SFP+/SFP28/QSFP+/QSFP56/QSFP28 connectors on both ends, which can provide an electrical connection directly into active equipment. DAC cables can be classified into twofold: passive DAC & active DAC. Both passive and active DAC cables can transmit electrical signals directly over copper cable. The former can deliver without signal conditioning, while the latter has electric components inside the transceivers to boost signals. Normally speaking, DAC cables are used to connect switches, servers, and storage inside racks. For DAC cables, please read Use High Speed Direct Attach Cable for Data Center Interconnection.

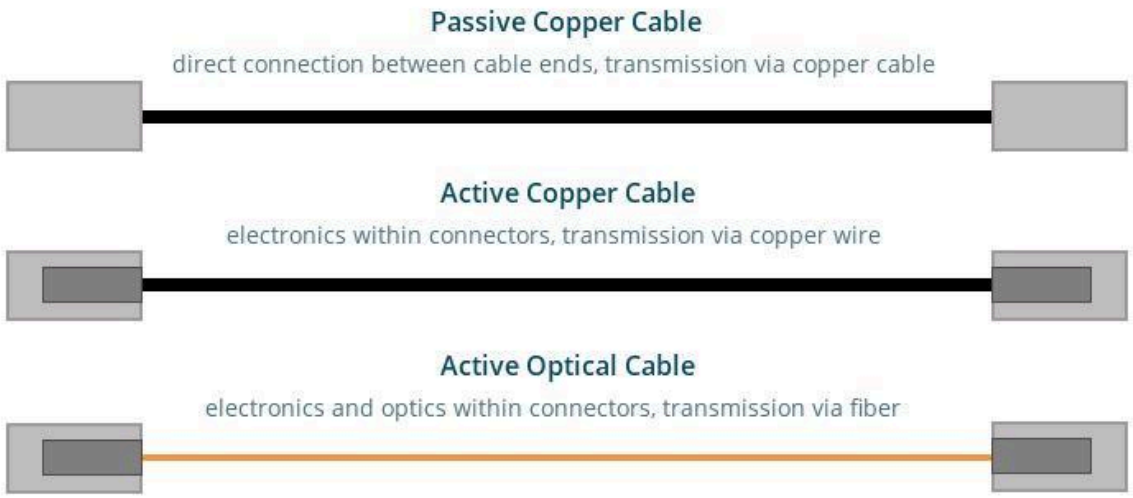


Figure 1: Passive DAC vs Active DAC vs AOC

Active Optical Cable (AOC) consists of a multimode fiber optic cable terminated with SFP form factor connectors on both ends, which requires external power to complete the conversion of electric and optical signals, from the electric signals to optical ones, and then convert to electric signals finally. Generally speaking, AOC cables are mostly used to link switches, servers, and storage between different racks inside data centers.

DAC/AOC Cables Comparison

DAC cables are used to connect switches, servers, and storage inside racks, while AOC cables are mostly used to link switches, servers, and storage between different racks inside data centers. Besides this, DAC AOC cables differ from each other in the following aspects.

	Reach	Cable Tyoes	Power Comsumption	Bend Radius	Application	Price
Passive DAC Cable	<7m	Twinax copper cable	<0.15w	24 AWG=38 mm 30 AWG=23 mm	ToR, Adjacent racks	$
Active DAC Cable	7-15m	Twinax copper cable	0.5-lw	24 AWG=38 mm 30AWG=23 mm	ToR, Adjacent racks	$$
Active Optical Cable	Up to 100m	Optical fiber	<1w	3.0mm	ToR EoR MoR	$$$

Power Consumption

Normally, the power consumption of AOC cables is higher than DAC ones, which is 1-2w. While the power consumption of DAC active cables is less than 1w, and the passive ones cost nearly no power consumption at the value of lower than 0.15w due to the thermal design of direct attach copper cables. As a result, the operating expenses on power consumption will be decreased when adopting the DAC options.

Transmission Distance

Adopting optical fiber technology, AOC cable can transmit over longer distances—100 m, while DAC cable link length limit is 10 m (passive DAC: 7 m; active DAC: 10 m). To sum up, DAC cabling solutions are suitable for short-range transmissions, while AOC solutions are applied in long-range networking cases.

Note: the max. distance of a signal that can be transmitted via a DAC cable change depending on the data rate. The link length will decrease as the data rate grows, for example, 100G DAC cables can only transmit up to 5 meters.

Cost

Roughly speaking, DAC has a relatively simple internal structure with fewer components, and the copper cables are much cheaper than fiber cables. When implemented in large-scale data centers, the sum of money will be saved for large quantities of DAC cables compared to AOC options. DAC indeed provides a cost-effective solution over AOCs for short-range applications, but for long-range applications, it's wise to have the overall costs list by comparing these two options.

EMI Immunity

Electromagnetic interference (EMI), refers to a disturbance generated by an external source that will affect the electrical circuit. Like mentioned before, the active optical cable contains optical fibers—a kind of dielectric that can't conduct electric current. Therefore, AOC cables are immune to electromagnetic interference, which can be used in most situations. However, due to the nature of copper with sending electrical signals, direct attach copper cables are vulnerable to the effects of EMI. Thus, the environment is important to avoid undesirable responses, degradation, or complete system failure.

DAC/AOC Working Scenarios

Influenced by the abovementioned factors, DAC and AOC cables are normally applied in different working scenarios.

DAC Cable Typical Application

The major utilization of 10G SFP+ DACs is connecting switches/servers to switches within or adjacent to the rack. In other words, these 10G direct attach cables can be used as an alternative for ToR (Top of Rack) interconnections between 10G ToR switch and server or the stacking of 10GbE switches. Since 10G SFP+DAC typically supports a link length of 7 m with low power consumption, low latency and low cost, this option is an ideal choice for this short-range server-to-switch connections.

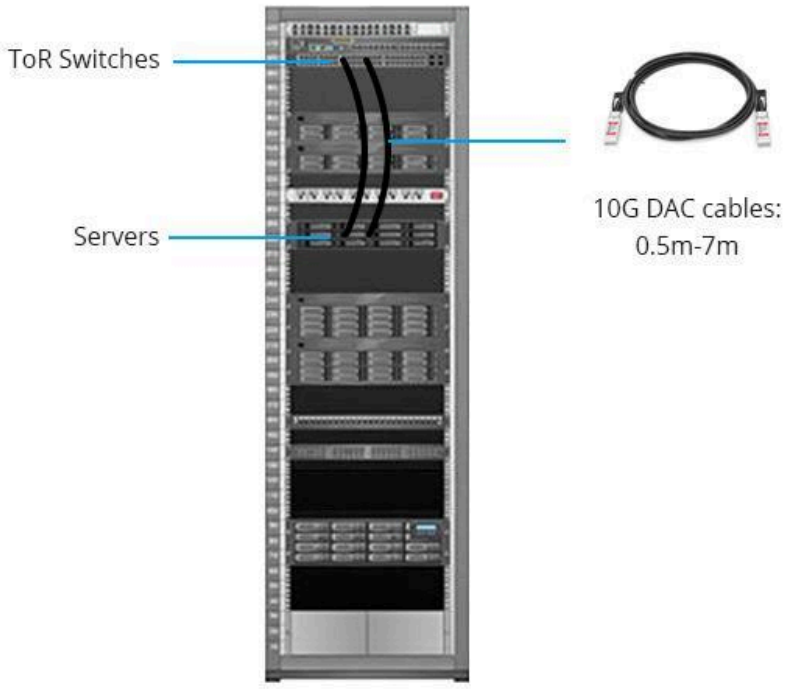


Figure 2: 10G DAC Connection Scenario

10G SFP+ AOC Cable Typical Application

Without the strict link length limits, 10G SFP+ AOCs are commonly used in several locations in the data center like ToR, EoR (End of Row) and MoR (Middle of Row). Like the DACs, the servers all connect up to a Top of Rack Ethernet switch, and each of them will have one or two Ethernet connections up to the switch and these can be patched by using AOC cables.

What's more, the utilization of 10G AOCs in the data center can also be realized in several main networking areas like Spine, Leaf or Core switching areas. The interconnections are typically fulfilled by adopting these 10G SFP+ AOCs with their theoretical maximum reach being 100 m.

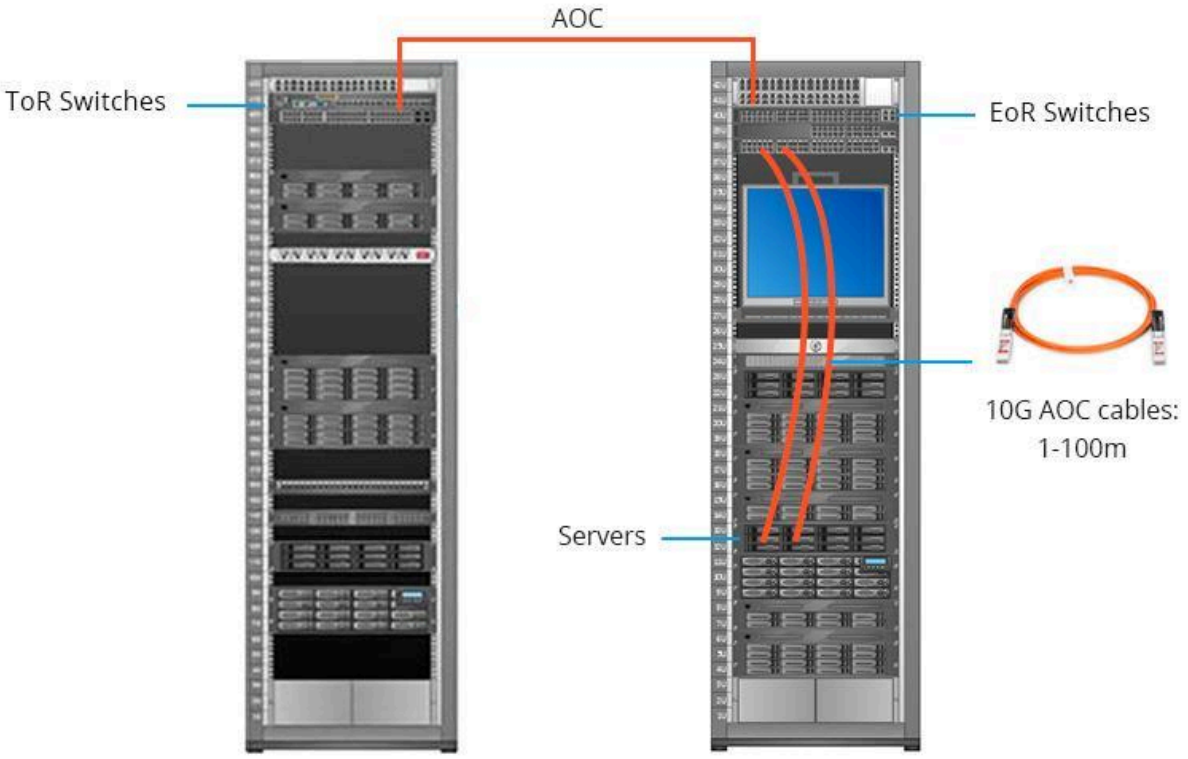


Figure 3: 10G AOC Connection Scenario

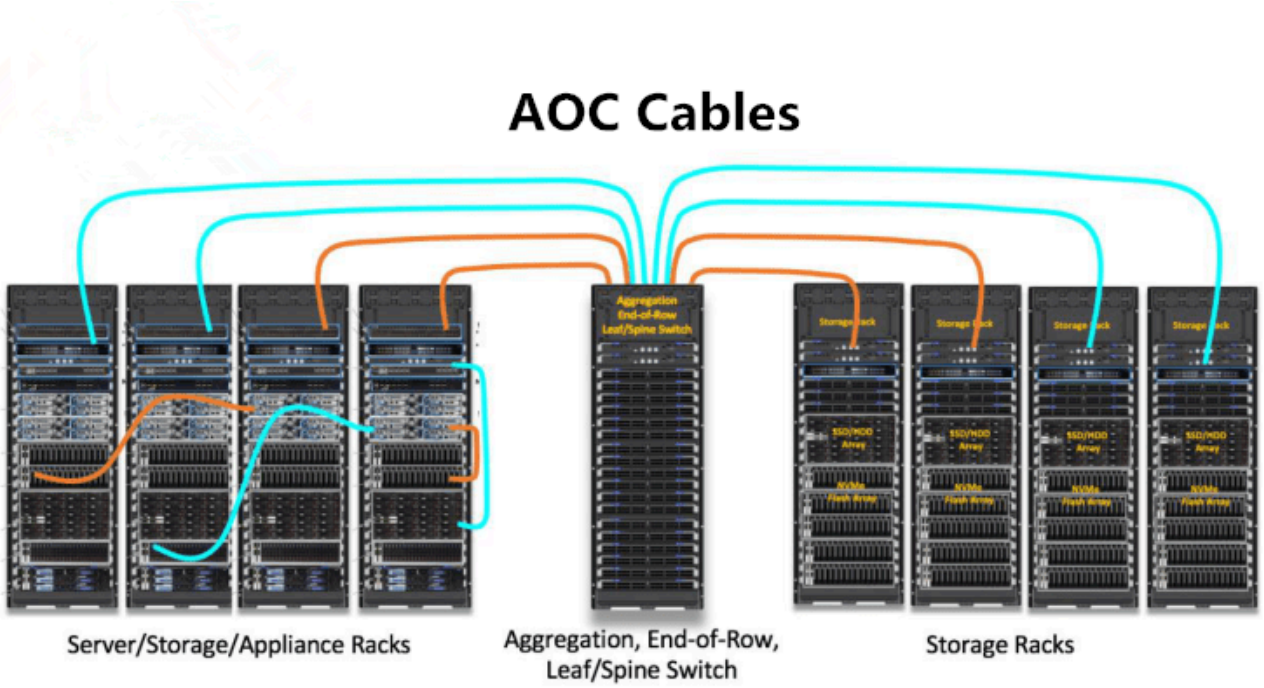
Active Optical Cable (AOC) – Rising Star of Telecommunications & Datacom Transceiver Markets

https://community.fs.com/blog/active-optical-cable-aoc-rising-star-of-telecommunications-datacom-transceiver-markets.html

Sheldon

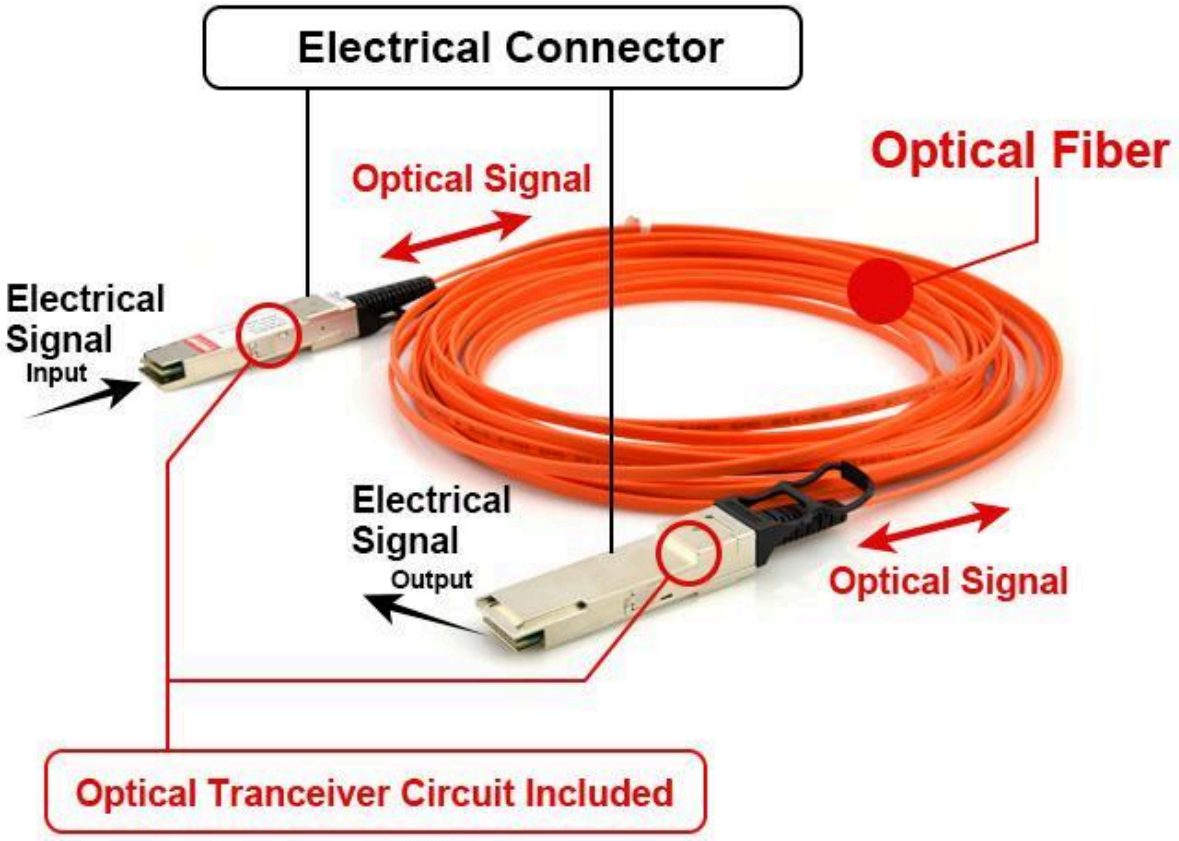
February 26, 2021

As data rates rise and data center clusters grow in size, copper cable technology is getting stretched to its limits. Active optical cable (AOC) assemblies were invented to replace copper technology in data centers and high performance computing (HPC) applications in virtue of its stability and flexibility. In the era of the optical network, AOC cable captures a major chunk of high-speed optical inter-connectivity market on a global level.



Active Optical Cable AOC Wiki

Active optical cable (AOC) can be defined as an optical fiber jumper cable terminated with optical transceivers on both ends. It uses electrical-to-optical conversion on the cable ends to improve speed and distance performance of the cable without sacrificing compatibility with standard electrical interfaces. To understand the structure of AOC cable, the following figure depicts the detailed structure:



Active Optical Cable Structure

Active optical cables incorporate active electrical and optical components to boost and receive signal via optical fiber. They are becoming one of the most popular cabling solutions in the data center. In the modern market, a variety of AOC cables have been launched for 10G/40G/100G applications, including 10G SFP+ AOC, 25G SFP28 AOC, 40G QSFP+ AOC, 56G QSFP+ AOC, 40G QSFP+ to 4x SFP+ breakout AOC, 40G QSFP+ to 8x LC breakout AOC, 100G QSFP28 AOC, 100G QSFP28 to 4x SFP28 breakout AOC and 120G CXP AOC, etc. These AOC optics are commonly used for short-range multi-lane data communication and interconnect applications between two devices, such as rack-to-rack, shelf-to-shelf interconnect, storage, hubs, switches, routers, servers, etc.

Why Use Active Optical Cable (AOC) Rather Than DAC or Fiber Transceiver?

For 10/40/100G transmission, there are many choices to meet the increasing needs for speed and performance. Why on earth do we need to choose active optical cable? The following part will demonstrate the potential advantages of AOC cable by comparing it with DAC cable and fiber optical transceiver.

	Direct Attach Copper (DAC)	Active Optical Cable (AOC)	2 Transceivers + Structured Cabling
Reach	< 15m	100m	Per TIA/IEEE Standard
Cable Type	Twinax Copper	Fiber Optic	Twisted pair copper or fiber optic
Power Comsumption	<1w	1-2w	1-2w
Weight	Heaviest Weight	Lightest Weight	Depending on structured cabling deployed
Bend Radius	24AWG=38mm 26AWG=33mm 28AWG=25mm 30AWG=23mm	25mm	Depending on structured cabling deployed
Application	Top of Rack (ToR) Adjacent racks	Top of Rack (ToR) Middle of Row End of Row Zone-to-zone	Top of Rack (ToR) Middle of Row End of Row Zone-to-zone
EMI	Weak	Strong	Strong
Price	$$	$$$	$$$$

From the chart, we can conclude that:

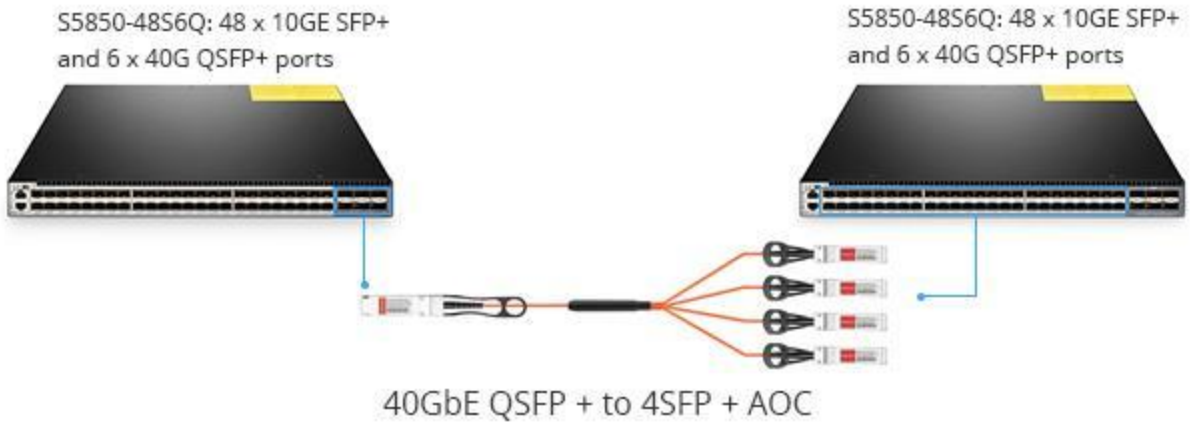
1. The lower weight and smaller bend radius of AOCs enable simpler cable management in high-density deployments.
2. The thinner AOC cables frees up a lot of space for increased air flow that is better helping to balance generated heat in crowded systems.
3. AOCs have great advantages over DACs especially when transmission distance reaches above 15 meters.
4. Optical fiber in AOC has considerable dielectricity, thus its EMI resistance level is considerably high.
5. AOC is a cost-effective and flexible short-reach (≤100m) direct-attach option in the interconnect system.
6. Since the connectors of AOC are factory pre-terminated, it is less affected by the repeating plug during daily use. It has also been proved that AOC has better reliability than that of transceivers.

How Are AOC Cables Used in Data Centers?

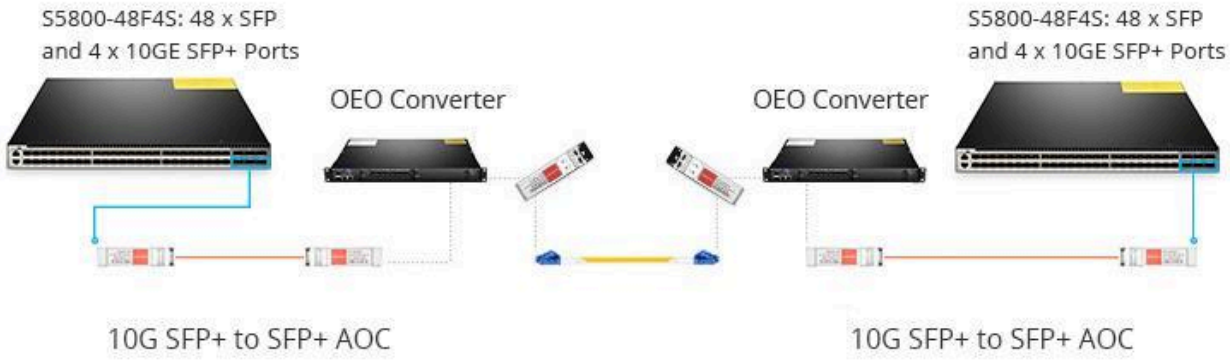
AOC cables are far superior in performance and their advantages are undeniable. When making plans for 10G/25G at the server and 40G/100G in the switching fabrics, multiple hyper-scale data center operators have completely shifted from using cooper cables to AOC cables. AOCs provide a cost-effective means to connect top-of-row (ToR) switches to end-of-row aggregation switches. Additionally, AOCs are also used to connect ToR switch with storage subsystems at reaches greater than DAC limits of 3-7 meters. The following three scenarios show the specific applications.

Scenario 1: One AOC cable can be used to connect two switches directly.

Scenario 2: The breakout AOC cable can offer a highly cost-effective way to connect within racks and across adjacent racks. The following figure shows 40GbE QSFP+ to 4 x SFP+ AOC cable connecting to a 40G QSFP+ port switch on one end, and to four 10G SFP+ ports switch on the other end.



Scenario 3: For the long haul transmission between the two switches, a suitable solution is suggested to use single-mode patch cable, OEO converters and AOC cables, which can provide seamless integration of different fiber types by converting multimode fiber to single-mode fiber.



What to Consider When Buying AOC cable?

Similar to buying other fiber optics, there are many details to consider when choosing AOC cables:

AOC Cable Data Rate

AOC cables are designed for data rates including 10Gb/s, 25Gb/s, 40Gb/s, 100Gb/s and 120Gb/s, which accelerate storage, data, and high-performance computing connectivity. In the AOC market, data center managers seem to appreciate the affordability of a plug-and-play, high-bandwidth 40G link that AOC offers.

AOC Cable Length

AOC cables are designed to bypass the bulk and distance limitations of copper. AOC cable assemblies are typically used in data centers for 1- 100m link length. As for how to make the final decision about the cable length, it depends on your specific situation. Too long or too short, it is susceptible to manage the cabling.

AOC Cable Reliability

Another issue to consider when choosing an AOC is reliability. As data rates increase and customers become less tolerant of errors and failure, the reliability of all equipment becomes more critical. The tiny electronics embedded in the transceiver carry a potential for failure. It is a wise action to choose an AOC vendor that can show testing confirming its product's reliability.

AOC Cable Price & Vendor

On the flip side, which brand should we choose for the AOC cables? Cisco, Juniper, HP, etc. are the most common brands, however products of these brands can only be applied in the corresponding switches, and the price of original pieces are super high. From an economic standpoint, many IT technicians are more inclined to a decent alternative to third-party optics.

AOC Market

While many brands of AOCs are available on the market, all are not alike. As data rates increase, concern over power consumption and heat generation will increase. Low-power AOC cable assemblies will become more and more important as data center operators strive to lower carbon footprint and energy costs.

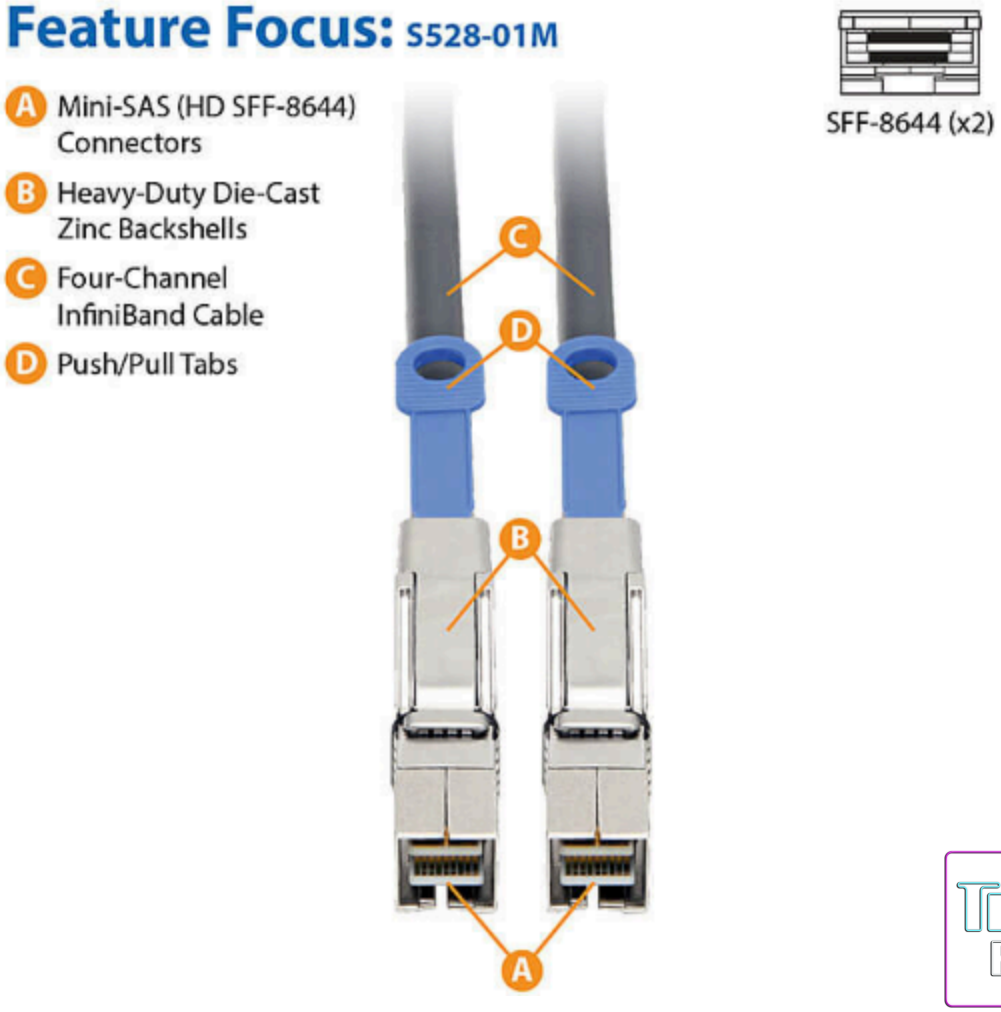
FS.COM Active Optical Cable Assemblies

FS.COM goes further than other third-party vendors, we use the same software codes as the original vendor for major-brands optics, and our generic transceivers can be compatible with

almost all the brands. All compatibility is strictly tested and controlled. More importantly, you can customize your AOC cables in accordance with your needs. The following offers FS.COM active optical cable assemblies solution:

Type	Description	Price
10G SFP+ AOC Cable	1m (3ft) Cisco SFP-10G-AOC1M Active Optical Cable	$ 42.00
25G SFP28 AOC Cable	1m Cisco SFP28-25G-AOC1M Active Optical Cable	$ 220.00
40G QSFP+ AOC Cable	1m Cisco QSFP-H40G-AOC1M Active Optical Cable	$ 120.00
40G QSFP+ to 4xSFP+ AOC Cable	1m Cisco QSFP-4X10G-AOC1M Active Optical Cable	$ 190.00
40G QSFP+ to 8xLC AOC Cable	5m (16ft) Cisco QSFP-8LC-AOC5M Active Optical Cable	$ 140.00
100G QSFP28 AOC Cable	1m Cisco QSFP-100G-AOC1M Active Optical Cable	$ 480.00
100G QSFP28 to 4x SFP28 AOC Cable	1m (3ft) Cisco QSFP-4SFP25G-AOC1M Active Optical Cable	$ 720.00

For more information about FS AOC cable, please contact us via sales@fs.com.
Related Article: Use High Speed Direct Attach Cable for Data Center Interconnection



Used for 3/6GB SAS external connections

EXTERNAL MINI-SAS CABLE (PULL-TAB) - 4X MINI-SAS (SFF-8088) TO 4X MINI-SAS (SFF-8088)

What is a SAN LUN?

by **Brian Reisdorf**, on Oct 31, 2019 7:44:37 AM

When dealing with the setup and configuration of Storage Area Networks (SANs), you may find yourself running into the term "LUN" more than a few times. Luckily, unless you're actually setting up the nuts and bolts portion of a new or reconfigured SAN, you shouldn't have to deal with LUNs very often. So what are they and what do they do?

What is a LUN?

LUN stands for "Logical Unit Number". It is a virtual address to a device in a SCSI environment. The short version here is that if you have a SCSI RAID device, each associated volume will have its own LUN assignment. This address tells the system which volume to send and read data from when it's addressed. In a single SCSI environment, no two devices can contain the same LUN or they will conflict, usually resulting in one of the devices becoming inaccessible or invisible on the SAN fabric until a unique LUN is assigned.

Why do we need LUNs?

As mentioned above, a LUN is an address in a SCSI environment that allows one system to find a data store, typically a disk partition, in a larger network. It's possible to have multiple LUNs on a single physical drive, all pointing at different disk partitions. By targeting and masking certain LUNs, we can dictate which systems are exposed to each other in our SAN. This level of control allows us to shape our SAN environment to fit our facilities needs.

Managing LUNs

LUNs allow a system architect to selectively reveal a device or partition to one or more servers or workstations on the SAN fabric. By configuring which LUNs

that a workstation or server can see on the network, you can restrict access to data or shape bandwidth paths based on zoned areas of the fabric and which LUNs those paths can reach. This allows for the configuration of redundant connection paths and paths with more bandwidth than others. Some SCSI based environments like Fibre Channel SAN system may have hundreds of individual LUNs associated with them. Most SAN software has some level of LUN management built into it, allowing you to adjust and refine what devices can see each other on the fabric.

Conclusion

LUNs are an integral part of the structure needed to properly deploy a SAN system. Set up properly, they allow us to control access and performance across a large environment of devices and storage. Once setup though, most users will not even know they exist, and given the somewhat complex nature of how they operate, that's not a bad thing.

Solid State Drive Form Factors

https://www.snia.org/forums/cmsi/knowledge/formfactors

Solid-state drives (SSDs) are commonly used in client, hyperscale and enterprise compute environments. They typically come in three flavors: NVMe™, SAS, and SATA. Since SSDs are made from flash memory, they can be built in many different form factors. This resource guide is designed to provide information on the most common and current SSDs in their various form factors. In addition to the form factor dimensions, information such as use case, interface, protocol, and mechanical/electrical and connector specifications are provided.

Click on the names below to learn more about the many different SSD sizes and formats in a variety of form factors:

- EDSFF
- M.2
- 2.5-inch (U.2)
- Add In Cards

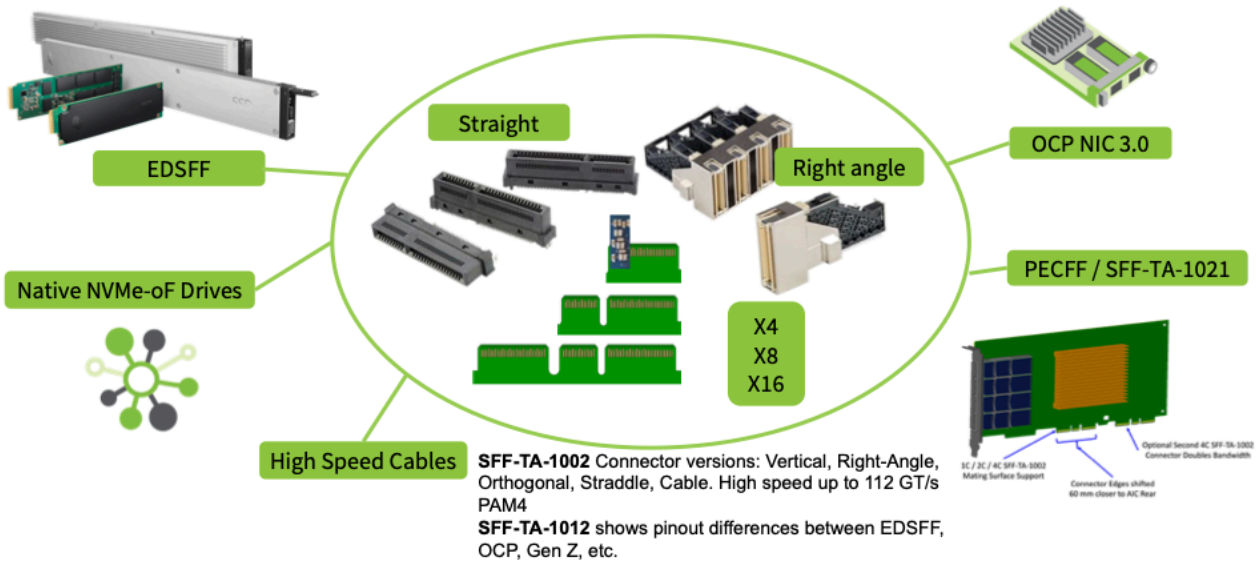
NVMe SSDs service many use cases in the data center. The NVMe SSD Classification page shows the different types of NVMe SSDs for different hyperscale and enterprise use cases.

The SNIA SFF Technology Affiliate is developing a broad range of standards for new connectors, form factors, and transceivers. **Learn more about their work**.

And for a lively debate on enterprise and data center standard form factors (EDSFF) - and if this is the end of the 2.5-inch disk era - watch this SNIA Video.

updated December 2021

EDSFF



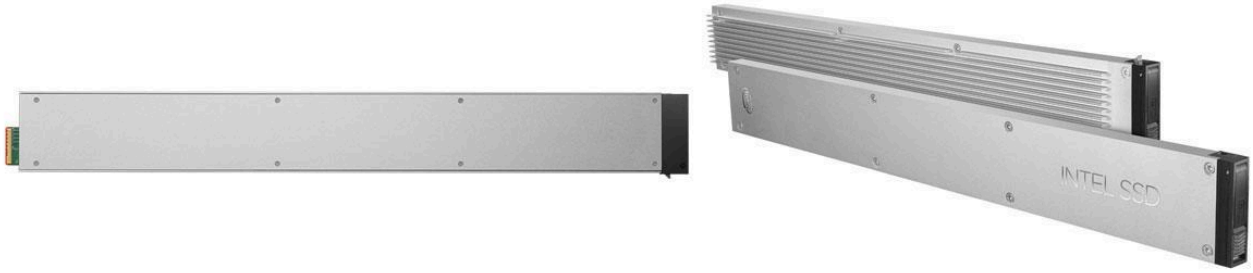
Click here for full image.

EDSFF stands for Enterprise and Data Center Standard Form Factor. The family of specifications were developed by a group of 15 companies working together to address the concerns of data center storage, and are now

maintained by SNIA as part of the SFF Technology Affiliate Technical Work Group (SFF TA TWG).

EDSFF offers a dynamic range of form factors that have advantages vs the incumbent SSD form factors in capacity, scalability, performance, serviceability, manageability, thermal and power management. Today all the EDSFF family of form factors share the same protocol (NVMe), the same interface (PCIe®), the same edge connector (SFF-TA-1002), the same pinout and functions (SFF-TA-1009). Infrastructure, especially test infrastructure, can be developed to support multiple EDSFF form factors. Learn more about the EDSFF family.

E1.L, EDSFF 1U Long



Illustrations left to right: E1.L 9.5mm (courtesy of Intel); E1.L 18mm (courtesy of Intel)

E1.L is a form factor that was developed to maximize capacity per drive and per rack unit in a 1U server or storage array (JBOD, JBOF), with superior manageability, serviceability, and thermal characteristics vs traditional form factors that were designed for rotating media. There are options for x4 or x8 lanes of PCIe while fitting vertically in a 1U chassis to allow for scalable bandwidth per drive, as well as options for 9.5 or 18mm heat sinks for various power and thermal environments (25W and 40W respectively). It improves data center serviceability, and is designed to be hot pluggable and front access serviceable with LEDs built into an integrated enclosure.

Use Cases

E1.L is optimized for high-capacity and dense storage use cases. High capacity per rack unit can improve data center TCO by offering storage consolidation and more power efficient storage (TB/W).

Dimensions

Type	Width	Length	Thickness
E1.L 9.5mm	up to 25W - 38.4mm	318.75mm	9.5mm
E1.L 18mm	up to 40W - 38.4mm	318.75mm	18mm

Mechanical/Electrical Specification:

- SFF-TA-1007 Version 1.2

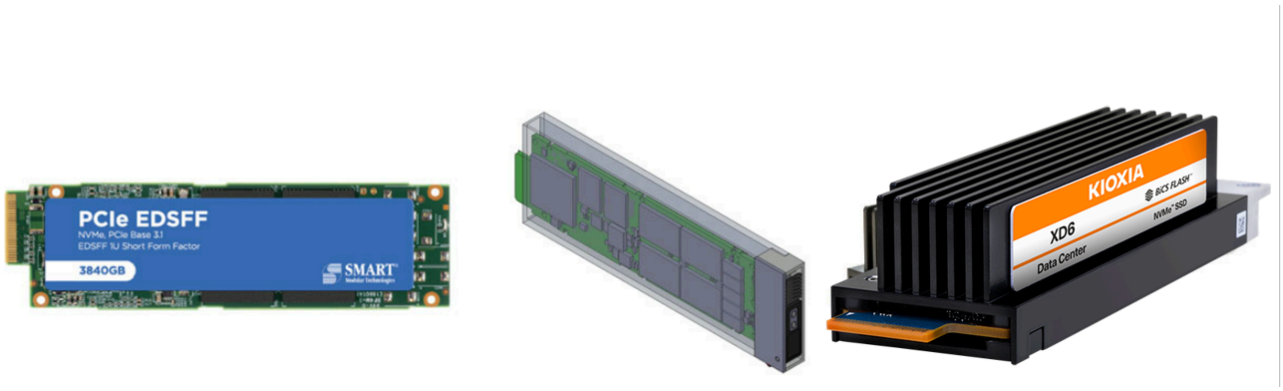
Connector Specification:

- SFF-TA-1002
- Pinouts SFF-TA-1009 (PCIe) and SFF-TA-1012 (others)

Protocol:

- NVMe over PCIe
- NVMe over Fabrics

E1.S, EDSFF 1U Short



Illustrations left to right: E1.S 5.9mm (courtesy of SMART Modular Systems); E1.S Symmetric Enclosure (courtesy of Intel); E1.S (courtesy of KIOXIA)

E1.S is a flexible, power efficient building block for hyperscale and enterprise compute nodes and storage. The M.2 110mm was popular in hyperscale data

centers due to the low-cost structure, flexibility, and scalability of multiple drives per server – but had challenges in hotplug / serviceability, thermals and overheating, and scaling to high capacities. E1.S solves those problems while maintaining the small form factor; E1.S is a small form factor being just a bit longer than M.2 but wider to accommodate more media (NAND) packages for increase capacity per drive. It fits vertically in a 1U chassis, similar to E1.L. The specification for E1.S 5.9mm has four standard mounting holes for heat sinks or carriers.

Different variants of E1.S offer improved flexibility for power, performance, scalability, and thermal efficiency. The latest version of E1.S offers a new optional symmetrical enclosure of 9.5mm width (similar to E1.L) that allows scalability up to 20W and x8 PCIe if required. Mainstream SSDs are still expected to be only PCIe x4, but the PCIe x8 support in the form factor allows use of other devices that need higher bandwidth.

The 15mm and 25mm asymmetrical enclosures offers a tradeoff of fewer drives per rack unit but improved power and performance per drive. At similar power per drive, the 15mm and 25mm enclosures offer improved cooling and thermal performance, decreasing the required airflow.

Use Cases

- Cloud compute servers
- OEM 1U performance server

Dimensions

Type	Width	Length	Thickness
E1.S 5.9mm	31.5mm	111.49mm	5.9mm
E1.S 8mm heat spreader	31.5mm	111.49mm	8.01mm
E1.S Symmetric Enclosure	33.75mm	118.75mm	9.5mm
E1.S Asymmetric Enclosure	33.75mm	118.75mm	15mm

E1.S Asymmetric Enclosure	33.75mm	118.75mm	25mm

Mechanical/Electrical Specification:

- SFF-TA-1006

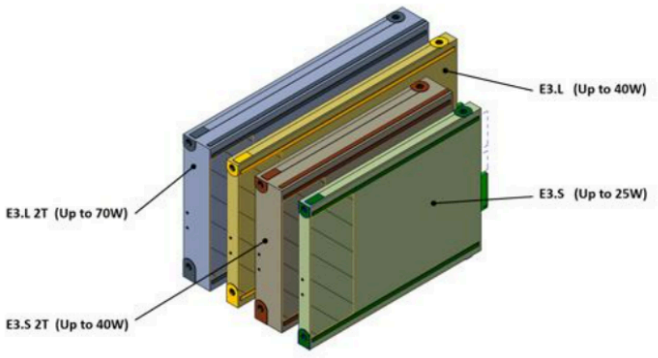
Connector Specification:

- SFF-TA-1002
- Pinouts SFF-TA-1009

Protocol:

- NVMe over PCIe
- NVMe over Fabrics

EDSFF E3.S and E3.L



Illustrations left to right: EDSFF E3 Group (SNIA specifications); E3.S (courtesy of KIOXIA)

EDSFF E3 is a family of form factors designed to update and replace the traditional U.2 2.5-inch form factor in servers and storage systems. The

different versions share the same height (76mm), and have commonality in length (112.75mm and 142.2mm) and width (7.5mm and 16.8mm). These hot-pluggable drives are designed more optimally for flash density of the SSD and system chassis. The E3 family connector is designed for x4 to x16 PCIe lanes and power envelopes up to 70W. All versions should be slot/connector compatible and are designed to be front accessible. Designed for future servers and storage systems, EDSFF E3 will accommodate next generations of PCI Express and can accommodate device types such as GPUs and NICs.

Where Used

- The primary usage is SSDs, but E3 is big enough to accommodate a broader range of device types.
- The E3 form factor allows for x4, x8, or x16 PCIe host interface.

Dimensions

Type	Height	Length	Width	Max Power
E3.S	76mm	112.75mm	7.5mm	25W
E3.S 2T	76mm	112.75mm	16.8mm	40W
E3.L	76mm	142.2mm	7.5mm	40W
E3.L 2T	76mm	142.2mm	16.8mm	70W

Mechanical/Electrical Specification:

- SFF-TA-1008 Rev 2.0
- SFF-TA-1023

Connector Specification:

- SFF-TA-1002
- Pinouts SFF-TA-1009 (PCIe) & SFF-TA-1012 (PCIe, Ethernet, Gen Z)

Protocol:

- NVMe over PCIe

- NVMe over Fabrics
- Native NVMe-oF Drive Specification

- Upcoming interfaces: Gen Z, CXL, OCP…

M.2

M.2 Illustrations left to right: courtely of KIOXIA, courtesy of Intel, courtesty of Samsung

M.2 is a form factor specification for internally mounted SSDs. Formerly known as Next Generation Form Factor (NGFF), M.2 supports PCIe, SATA and USB interfaces and comes in various widths and lengths. It also has keying notches on the edge connector to designate various interface or PCIe lane configurations. M.2 is smaller than the typical 2.5" form factor SSD and is typically removable, except Type 1620 (BGA), which offers a ball grid array chip package and it typically mounted on the main system board.

Dimensions

To view a visual representation of M.2 form factor dimensions, click here.

Name	Description	Use Cases	Dimensions (W x L x H, mm)

16mm x 20 mm	x2 or x4 lanes PCIe running the NVMe command set, surface mounted or on 22x30 M.2 PCB. Capacities up to 1TB	Mobile/Tablet/Laptop	16.15 x 20.15 x 1.3
22mm x 30mm	x2 or x4 lanes PCIe running the NVMe command set. May be BGA mounted on M.2. Capacities up to 1TB.	Mobile/Tablet/ Laptop/PC boot/ Server boot	22.15 x 30.15 x 2.23
22mm x 80mm	x4 lanes PCIe running the NVMe command set. Capacities up to 4TB.	Laptop/PC/ Server or Hyperscale data/ Server boot	22.15 x 80.15 x 2.23
22mm x 110mm	x4 lanes PCIe running the NVMe command set. Capacities up to 8TB.	PC boot and Data/ Server or Hyperscale data	22.15 x 110.15 x 3.88

Mechanical/Electrical Specification:

- PCI Express M.2 Specification Version 1.2, Revision 3.0
- SATA Specification Version 3.2

Connector Specification:

- PCI Express M.2 Specification Version 1.2, Revision 3.0
- SATA Specification Version 3.2

Protocol:

- SATA
- NVMe/PCIe

2.5-inch (U.2)

The 2.5-inch form factor is the most common deployment of an SSD, and is offered with PCIe (with NVMe), SAS or SATA interfaces. It is typically used in desktops, servers and storage systems built around hard disk drives (HDD). This form factor is commonly associated with the term U.2 and is sometimes referred to as the U.2 form factor. U.2 is defined as compliance with the PCI Express SFF-8639 Module specification, and no longer typically references SAS or SATA SSDs.

Name	Description	Use Cases	Dimensions (W x L x H, mm)
2.5-inch (7 mm)	Typically x4 NVMe, slim form factor, capacities up to 3.84TB	NVMe interface, PC or Server, Hyperscale environments	69.85 x 100 x 7
2.5-inch (7 mm)	Typically 6Gb/s SATA, slim form factor, capacities up to 3.84TB	SATA interface, PC or Server, Hyperscale environments	69.85 x 100 x 7
2.5-inch (15 mm)	Typically x4 NVMe, dual-port	NVMe interface, Server or Storage,	69.85 x 100 x 15

	support, capacities up to 30.72TB	Enterprise environments	
2.5-inch (15 mm)	Typically 12Gb/s SAS, dual-port support, capacities up to 30.72TB	SAS interface, Server or Storage, Enterprise environments	69.85 x 100 x 15

Mechanical/Electrical Specification:

- 2.5-inch (7 mm) - NVMe – PCI Express SFF-8639 Module
- 2.5-inch (15 mm) - NVMe-PCI Express SFF 8639 Module
- SFF 8201 2.5” Form Factor Drive Dimensions
- Enterprise SSD Form Factor V1.0a

Connector Specification:

- 2.5-inch (7 mm) -SFF-8639 Multifunction 6X Unshielded Connector
- 2.5-inch (15 mm) – SFF-8639 (all interfaces) SFF-8680 – SAS

Protocol:

- NVM Express Base Specification
- SATA
- SAS

Add-In Card (AIC)

An Add-in Card (AIC) is a solid-state device that utilizes a standard card form factor such as a PCIe card. The AIC would usually use an interface such as PCIe, or possibly a mezzanine card of a standard form type. Given the larger physical size over an SSM, the AIC would typically have larger capacity and potentially higher performance. In addition, the larger size allows for the potential to add computational function to the storage device. Because of the versatility of the form factor, the AIC is a form factor that likely will constantly evolve for solid-state storage.

Name	Description	Use Cases	Dimensions (in/mm)
Full Height	PCIe	Enterprise and cloud deployments, server deployments, large capacity, additional processing for security, storage functionality, and/or future expansion. Support of higher range of power options.	6.6 in/167 mm
Half Height	PCIe	Enterprise and cloud deployments, server deployments, large capacity, additional processing for security, storage functionality, and/or future expansion. Support of higher range of power options.	4.3 in/ 111 mm

Low Profile	PCIe	Enterprise and cloud deployments, server deployments, large capacity, additional processing for security, storage functionality, and/or future expansion. Support of higher range of power options.	2.5 in/64 mm
Full Length	PCIe	Enterprise and cloud deployments, server deployments, large capacity, additional processing for security, storage functionality, and/or future expansion. Support of higher range of power options.	12.2 in/312 mm
Half Length	PCIe	Enterprise and cloud deployments, server deployments, large capacity, additional processing for security, storage functionality, and/or future expansion. Support of higher range of power options.	6.6 in/167 mm

Actual SSC dimensions may be less, dependent on design.

PCIe cards can also come in multiple widths. Such a configuration allows mating with adjacent motherboard connectors, enabling increased performance by supporting more than 16 PCIe lanes.

https://youtu.be/raeUiNtMk0E Introduction to Microsoft's Storage Spaces Direct

Storage Spaces Direct overview

Applies to: Windows Server 2022, Windows Server 2019, Windows Server 2016

Storage Spaces Direct overview | Microsoft Docs

Storage Spaces Direct uses industry-standard servers with local-attached drives to create highly available, highly scalable software-defined storage at a fraction of the cost of traditional SAN or NAS arrays. Its converged or hyper-converged architecture radically simplifies procurement and deployment, while features such as caching, storage tiers, and erasure coding, together with the latest hardware innovations such as RDMA networking and NVMe drives, deliver unrivaled efficiency and performance.

Storage Spaces Direct is included in Windows Server 2019 Datacenter, Windows Server 2016 Datacenter, and Windows Server Insider Preview Builds. It also provides the software-defined storage layer for Azure Stack HCI.

For other applications of Storage Spaces, such as shared SAS clusters and stand-alone servers, see Storage Spaces overview. If you're looking for info about using Storage Spaces on a Windows 10 PC, see Storage Spaces in Windows 10.

TABLE 1

Description	Documentation
Understand • Overview (you are here) • Understand the cache	**Plan** • Hardware requirements • Using the CSV in-memory read cache • Choose drives

TABLE 1

Description	Documentation
• Fault tolerance and storage efficiency • Drive symmetry considerations • Understand and monitor storage resync • Understanding cluster and pool quorum • Cluster sets	• Plan volumes • Using guest VM clusters • Disaster recovery
Deploy • Deploy Storage Spaces Direct • Create volumes • Nested resiliency • Configure quorum • Upgrade a Storage Spaces Direct cluster to Windows Server 2019 • Understand and deploy persistent memory	**Manage** • Manage with Windows Admin Center • Add servers or drives • Taking a server offline for maintenance • Remove servers • Extend volumes • Delete volumes • Update drive firmware • Performance history • Delimit the allocation of volumes • Use Azure Monitor on a hyper-converged cluster
Troubleshooting • Troubleshooting scenarios • Troubleshoot health and operational states • Collect diagnostic data with Storage Spaces Direct • Storage-class memory health management	**Recent blog posts** • 13.7 million IOPS with Storage Spaces Direct: the new industry record for hyper-converged infrastructure • Hyper-converged infrastructure in Windows Server 2019 - the countdown clock starts now! • Five big announcements from the Windows Server Summit • 10,000 Storage Spaces Direct clusters and counting...

NEW! PCIe Gen 4 24G SAS/SATA/NVMe SmartRAID Adapters and Host Bus Adapters (HBAs)

We power the industry's first PCIe Gen 4 24G SAS/SATA/NVMe™ RAID adapters and HBAs with our fifth generation SmartROC 3200 and SmartIOC 2200 controllers. These RAID adapters and HBAs bring market-leading performance, industry-first capabilities and features such as high-performing NVMe RAID DirectPath technology for low-latency NVMe transactions and Dynamic Channel Multiplexing (DCM) for efficient aggregation. We provide endless design flexibility with the ability to operate NVMe, SAS and SATA storage devices in a single bay and we offer the widest breadth of qualified 24G NVMe/SAS/SATA PCIe board-level solutions, including a 32-port variant.

All critical data center security needs are delivered with hardware root of trust, Controller-Based Encryption (CBE) and Self-Encrypting Drive (SED) support. New management tools such as support for intelligent backplane management and PLDM/RDE to simplify the implementation of out-of-band management over MCTP/BMC and our common management experience across the Smart Storage platform and tri-mode media allows for faster development and deployment.

Adaptec® SmartRAID 3200 RAID Adapters new 24G SAS controllers

SmartRAID 3200 adapters deliver a full RAID feature set that includes:

- The industry's first 24G SAS adapter support for industry-standard SAS-4 connectivity
- Support for both ×8 and ×16 PCIe Gen 4 host interface adapters with 32, 16 and 8 media-facing port variants for internal and external tri-mode connectivity
- Integrated PCIe switch that enables DirectPath technology for the industry's lowest latency and highest bandwidth NVMe solution with the flexibility to support ×1-, ×2-, ×4- and ×8-wide NVMe SSDs
- Best-in-class ultra-performance adapters with a ×16 host interface which doubles the bandwidth to 29.6 GB/s (four times that of the previous generation)
- Ultra-performance adapters support up to 11.7 GB/s throughput on RAID 5 redundant writes (45% higher than 72-bit products) and 17 Gbps throughput on NVMe CBE-enabled logical devices using maxCrypto™ encryption
- Our **Dynamic Channel Multiplexing (DCM)** technology aggregates expander-attached SAS or SATA hard drives onto 24G SAS infrastructure with near 100% link efficiency for unparalleled throughput that supports all media types
- Support for Intel VPP intelligent backplane management and UBM standards to simplify integration and enhance product flexibility for system integrators

Download SmartRAID 3200 Sell Sheet

Adaptec SmartHBA 2200 and HBA 1200 Host Bus Adapters (HBAs)

SmartHBA 2200 and HBA 1200 adapters deliver a comprehensive feature set including:

- Widest breadth of qualified 24G NVMe/SAS/SATA PCIe board-level solutions, including an HBA 1200 32-port variant
- The industry's first 24G SAS HBA support for industry-standard SAS-4 connectivity
- Combination of full-featured, high-performance HBA functionality with basic RAID hardware with the SmartHBA 2200
- Superior performance enabling up to 29.6 GB/s throughput and 3.5M+ IOPs and 4K Read Writes (RR)
- Our DCM technology aggregates expander-attached SAS or SATA hard drives onto 24G SAS infrastructure with near 100% link efficiency for unparalleled throughput that supports all media types

- Integrated PCIe switch that enables DirectPath technology for the industry's lowest latency and highest bandwidth NVMe solution with the flexibility to support ×1, ×2, ×4 and ×8-wide NVMe SSDs
- Support for Intel VPP intelligent backplane management and UBM standards to simplify integration and enhance product flexibility for system integrators

Download SmartHBA 2200 Sell Sheet
Download HBA 1200 Sell Sheet